HDR+, portrait mode, Super Res Zoom, Night Sight:
computational photography and machine learning on Google's smartphones

Google AI Connect
San Francisco, February 1, 2019

Marc Levoy
Distinguished Engineer
Google Research

Professor, Emeritus
Computer Science Department
Stanford University

Trends in cell phone cameras

- there are billions of them, and still growing fast
- better pixels, better lenses, larger apertures
- multiple cameras
- depth sensors

Trends in cell phone camera <u>systems</u>

- software-defined camera
  - moving away from fixed-function hardware
  - combine bursts of frames (computational photography)
- machine learning
  - replacing classical algorithms for many tasks
  - more training data = better accuracy on these tasks
- less secrecy, more publication
  - forces faster innovation
  - attracts PhD superstars

The elephant in the room

"I'm right there in the room, and no one even acknowledges me."

- software and machine learning require more computation
- mobile devices are thermally constrained
- seldom better to send images to the cloud for analysis

Trends in cell phone camera systems

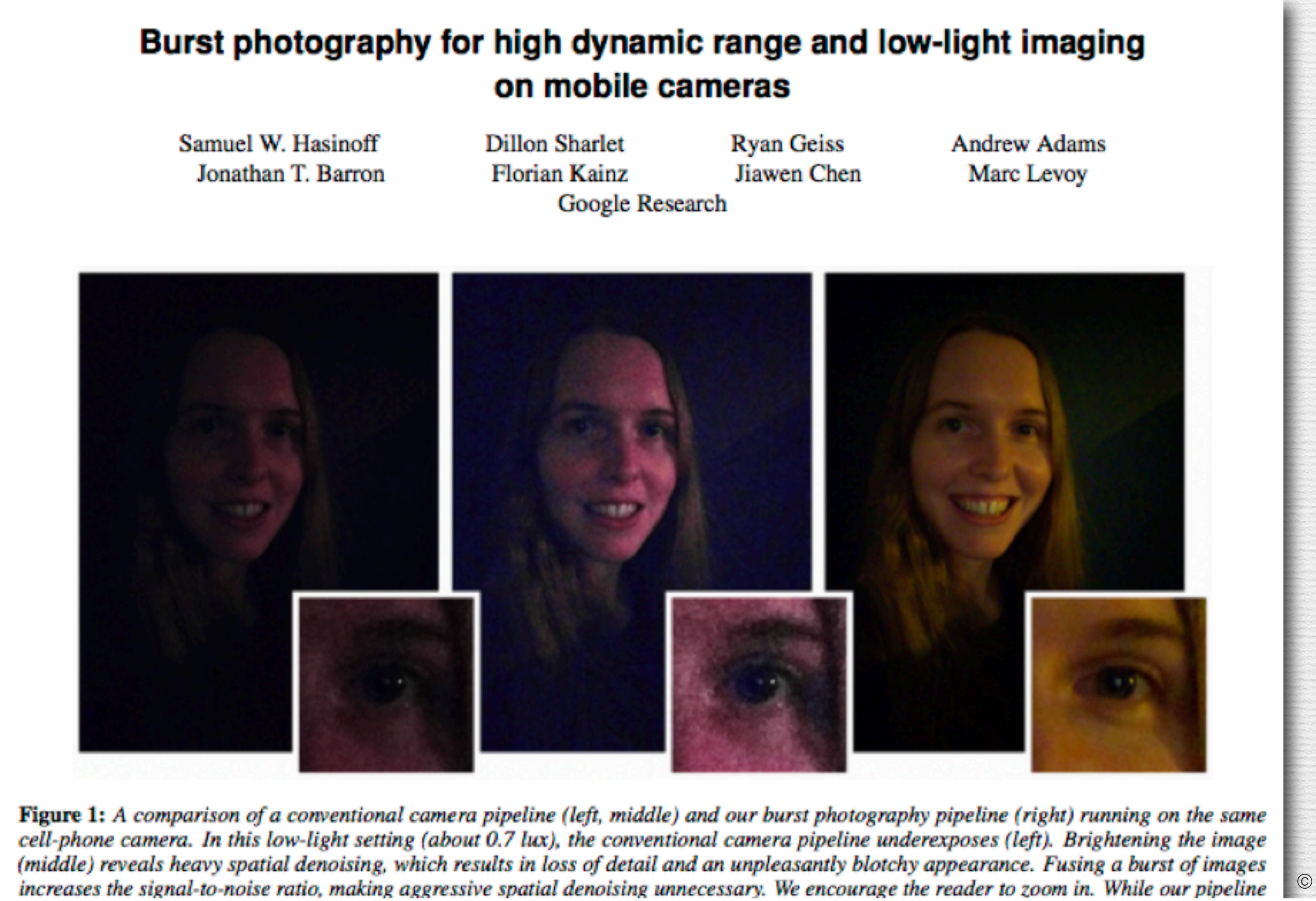
- software-defined camera
  - moving away from fixed-function hardware
  - combine bursts of frames (computational photography)
- machine learning
  - replacing classical algorithms for many tasks
  - more training data = better accuracy on these tasks
- less secrecy, more publication
  - forces faster innovation
  - attracts PhD superstars
- programmable hardware accelerators
  - CPUs, GPUs, DSPs (Pixel Visual Core)
  - Halide, GPGPU languages (CUDA, OpenCL, Vulkan)

Rules for cell phone camera apps

- ✦ the camera must feel fast
  - live viewfinder must be > 15fps
  - shutter lag must be < 150ms
  - photos must be ready in < 4 seconds
- ✦ default mode must never fail
  - reliable exposure, focus, and white balance
  - no ghosts or other visual artifacts, <u>ever</u>
- ✦ consumer photography is all about corner cases
  - after all, we look for “unusual photographs”
- ✦ occasional failures in special modes are ok
  - especially if they’re humorous

Example #1: HDR+ on Nexus/Pixel phones

[Hasinoff et al., SIGGRAPH Asia 2016]



Burst photography for high dynamic range and low-light imaging on mobile cameras

Samuel W. Hasinoff Dillon Sharlet Ryan Geiss Andrew Adams
Jonathan T. Barron Florian Kainz Jiawen Chen Marc Levoy
Google Research

Figure 1: *A comparison of a conventional camera pipeline (left, middle) and our burst photography pipeline (right) running on the same cell-phone camera. In this low-light setting (about 0.7 lux), the conventional camera pipeline underexposes (left). Brightening the image (middle) reveals heavy spatial denoising, which results in loss of detail and an unpleasantly blotchy appearance. Fusing a burst of images increases the signal-to-noise ratio, making aggressive spatial denoising unnecessary. We encourage the reader to zoom in. While our pipeline*

Typical approach to HDR

- exposure bracketing
  - capture images with varying exposure
  - combine highlights from short exposure with shadows from long exposure

- hard to robustly align images with camera shake or object motion
  - noise level differs between exposures
  - saturated areas cannot be aligned at all

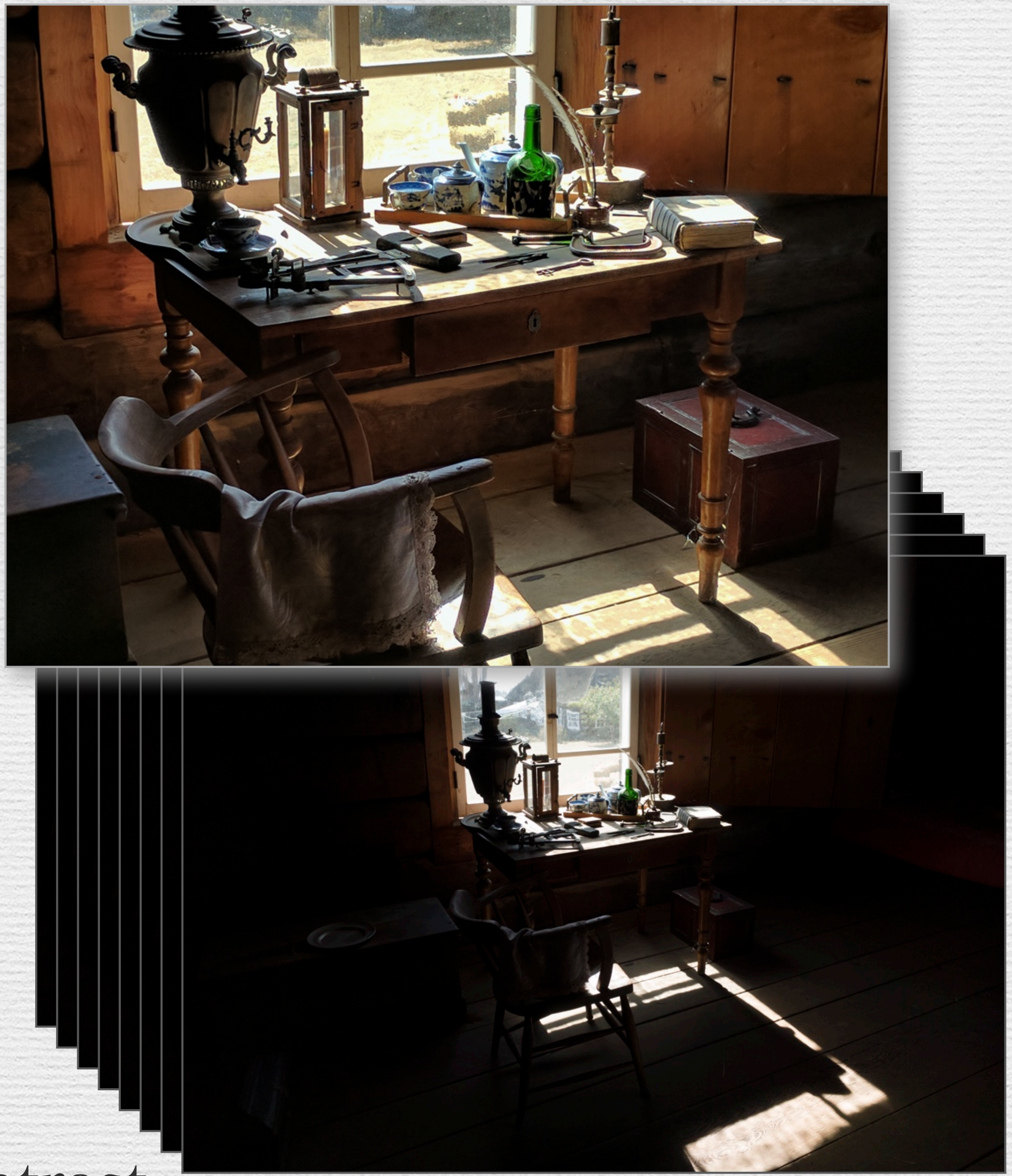
HDR+ in the Google camera app

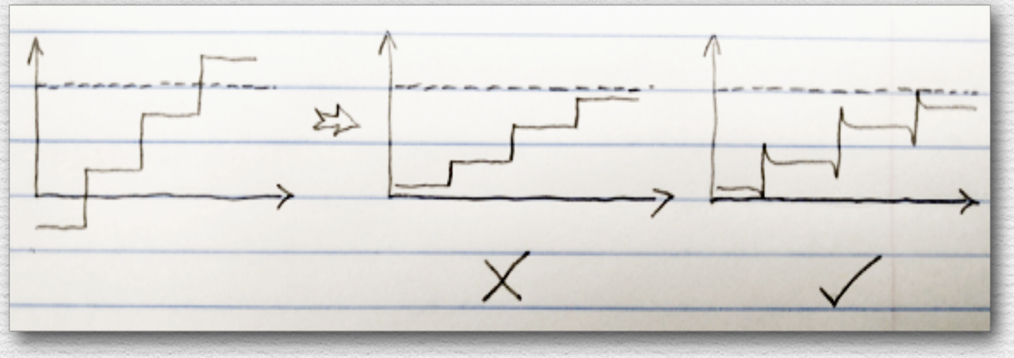
- capture a burst of under-exposed images
  - same exposure on all images in burst
  - avoids blowing out highlights

- align and merge
  - similar images align well
  - SNR ∝ sqrt(size of burst)
  - reduces noise in shadows

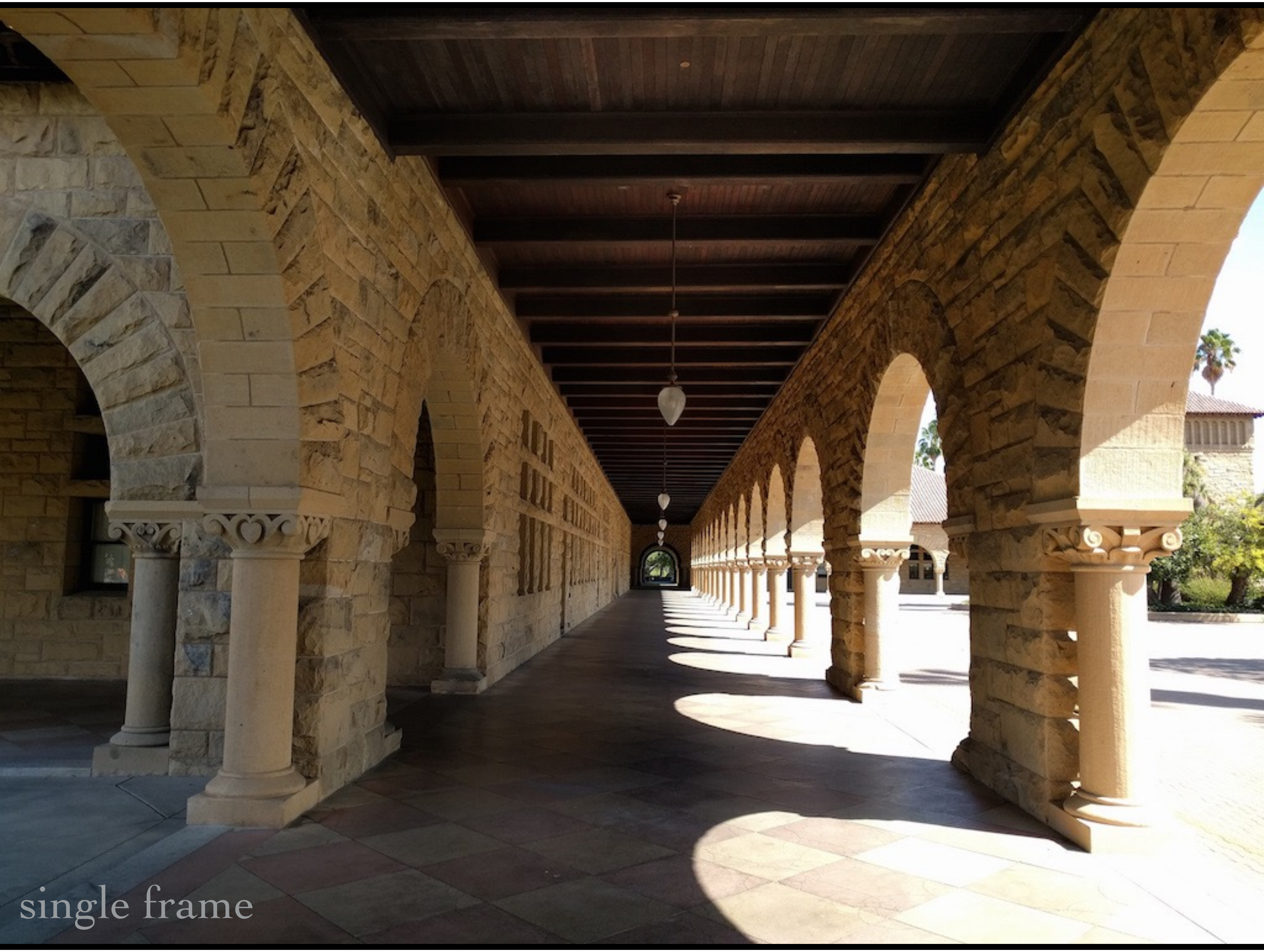
- tonemap
  - boost shadows
  - preserve local contrast at the expense of global contrast

HDR+ in the Google camera app

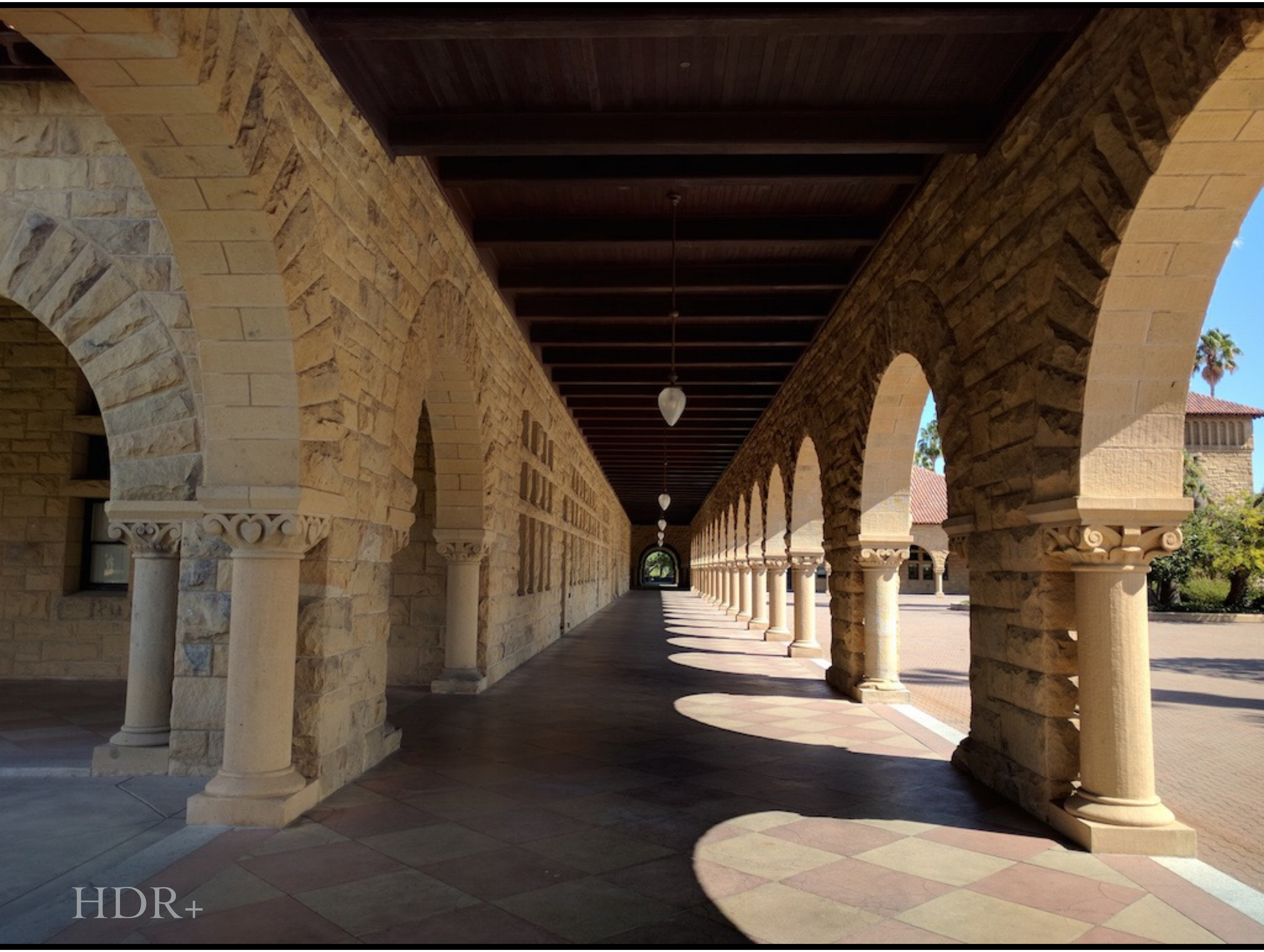




- ✦ tonemap
  - boost shadows
  - preserve local contrast
    at the expense of global contrast

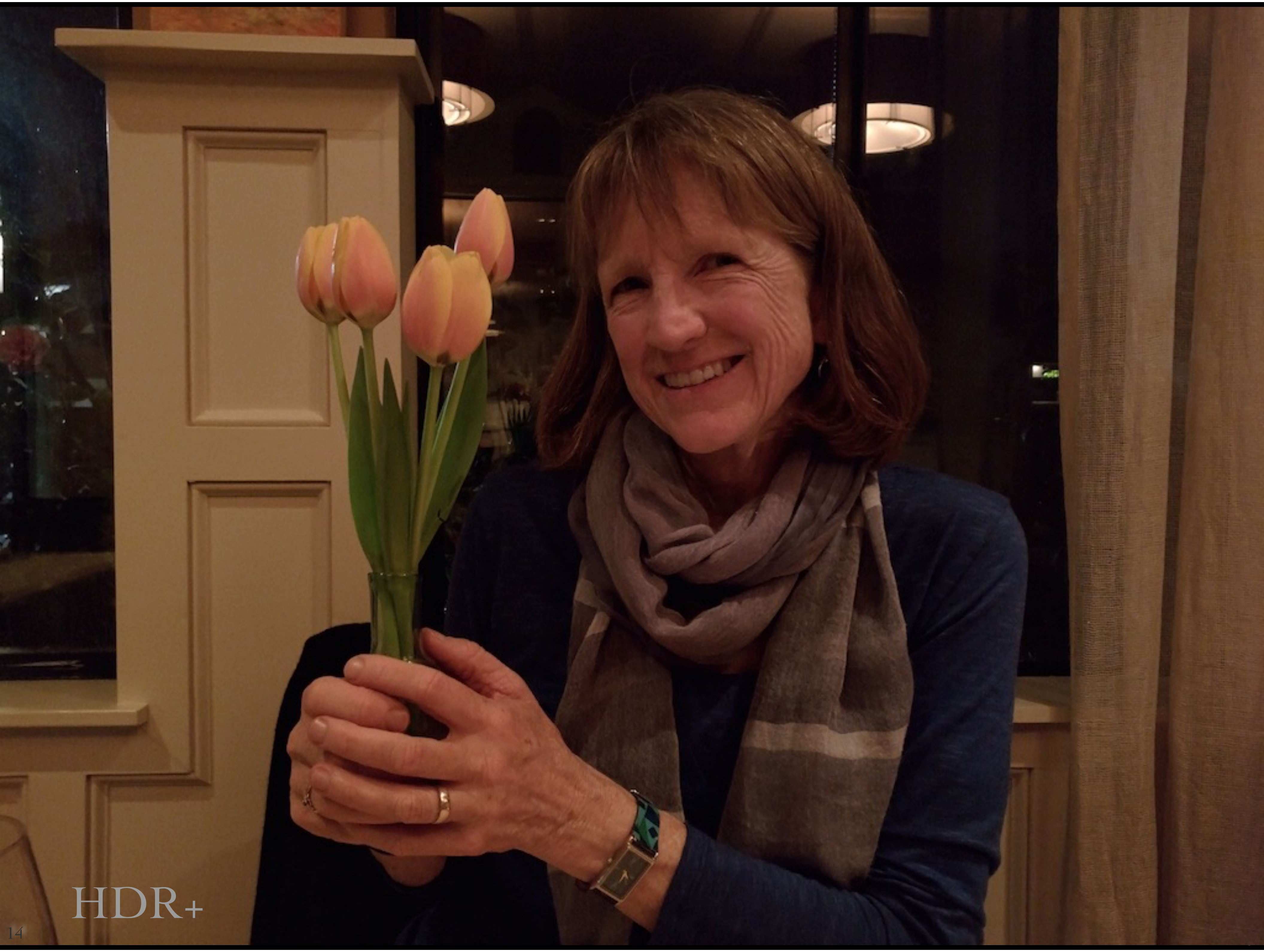


single frame



HDR+

single frame



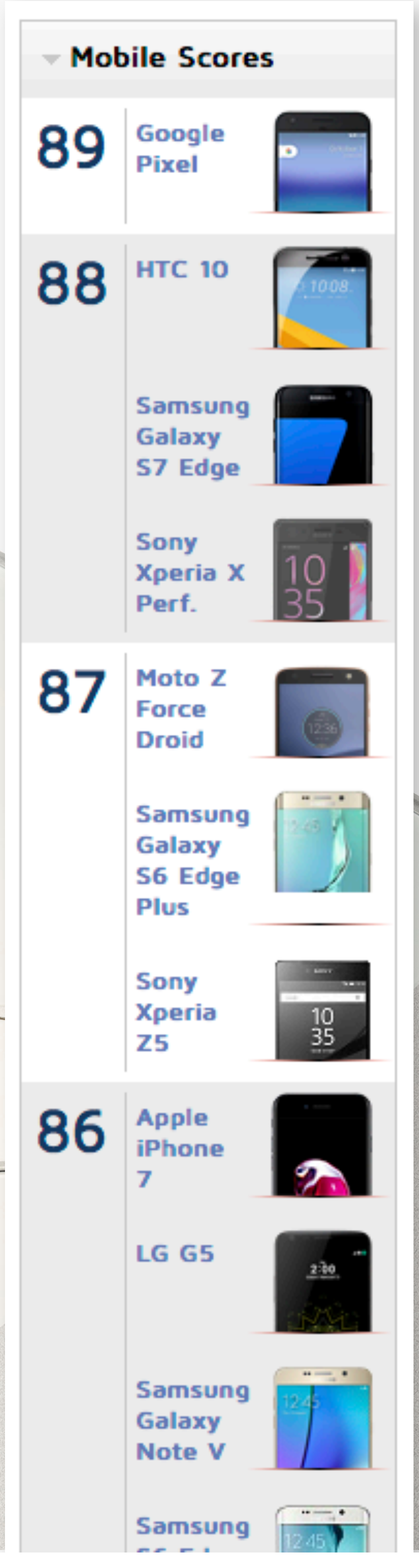
HDR+

Pixel

Phone by Google

DXOMARK MOBILE

89

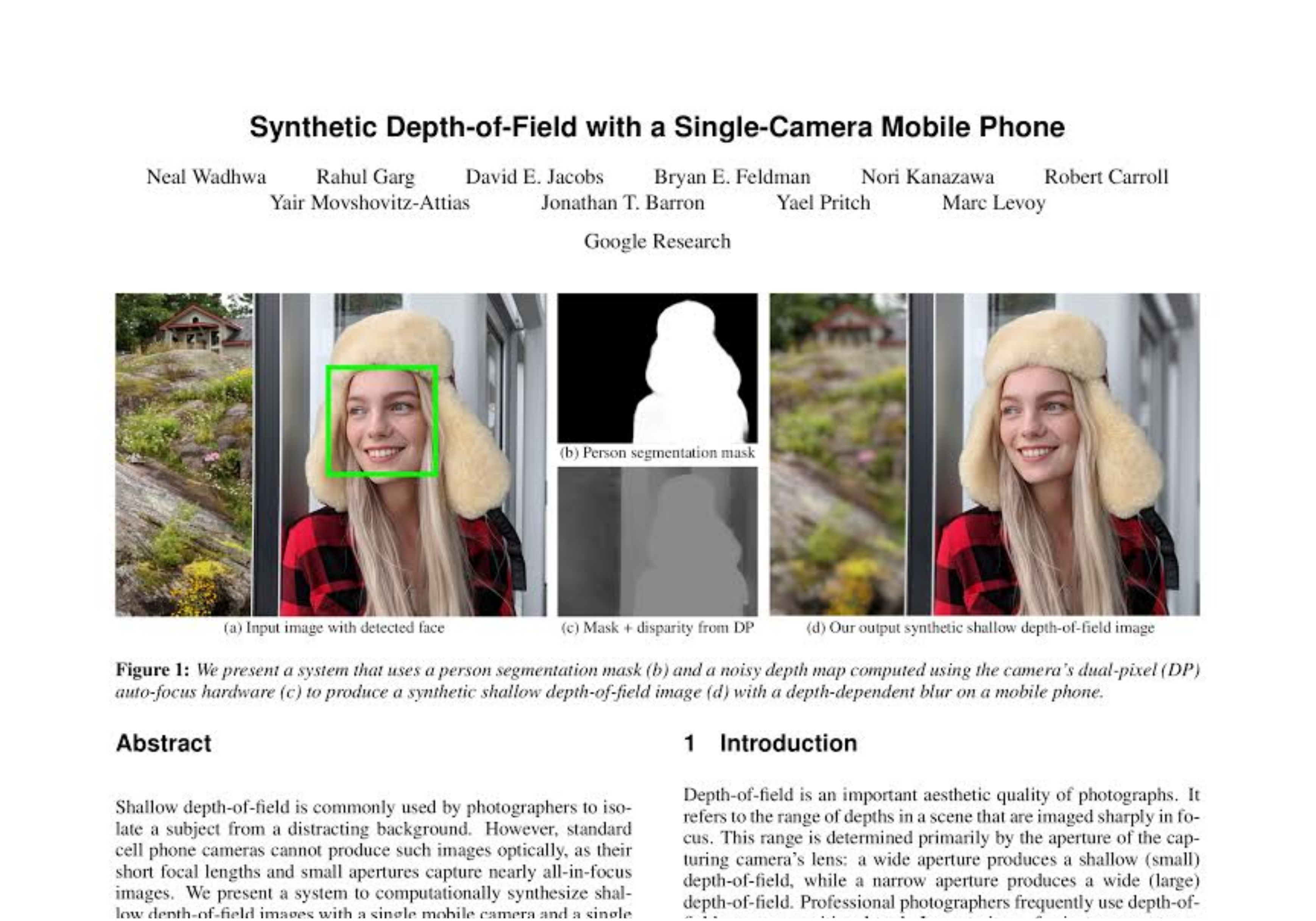


Mobile Scores

89	Google Pixel
88	HTC 10
	Samsung Galaxy S7 Edge
	Sony Xperia X Perf.
87	Moto Z Force Droid
	Samsung Galaxy S6 Edge Plus
	Sony Xperia Z5
86	Apple iPhone 7
	LG G5
	Samsung Galaxy Note V
	Samsung

Example #2: Portrait mode on the Pixel 2

[Wadhwa et al., SIGGRAPH 2018]



Synthetic Depth-of-Field with a Single-Camera Mobile Phone

Neal Wadhwa Rahul Garg David E. Jacobs Bryan E. Feldman Nori Kanazawa Robert Carroll
Yair Movshovitz-Attias Jonathan T. Barron Yael Pritch Marc Levoy

Google Research

(a) Input image with detected face

(b) Person segmentation mask

(c) Mask + disparity from DP

(d) Our output synthetic shallow depth-of-field image

Figure 1: *We present a system that uses a person segmentation mask (b) and a noisy depth map computed using the camera's dual-pixel (DP) auto-focus hardware (c) to produce a synthetic shallow depth-of-field image (d) with a depth-dependent blur on a mobile phone.*

Abstract

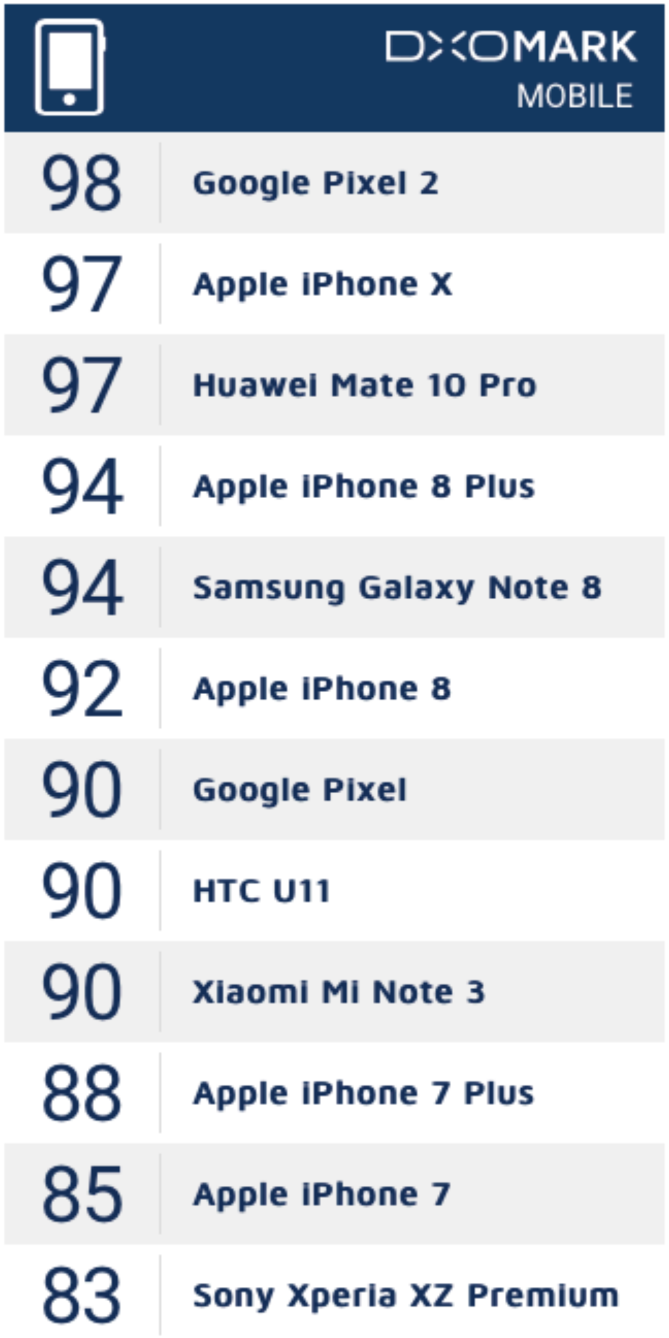
Shallow depth-of-field is commonly used by photographers to isolate a subject from a distracting background. However, standard cell phone cameras cannot produce such images optically, as their short focal lengths and small apertures capture nearly all-in-focus images. We present a system to computationally synthesize shallow depth-of-field images with a single mobile camera and a single

1 Introduction

Depth-of-field is an important aesthetic quality of photographs. It refers to the range of depths in a scene that are imaged sharply in focus. This range is determined primarily by the aperture of the capturing camera's lens: a wide aperture produces a shallow (small) depth-of-field, while a narrow aperture produces a wide (large) depth-of-field. Professional photographers frequently use depth-of-

LGI

LGI

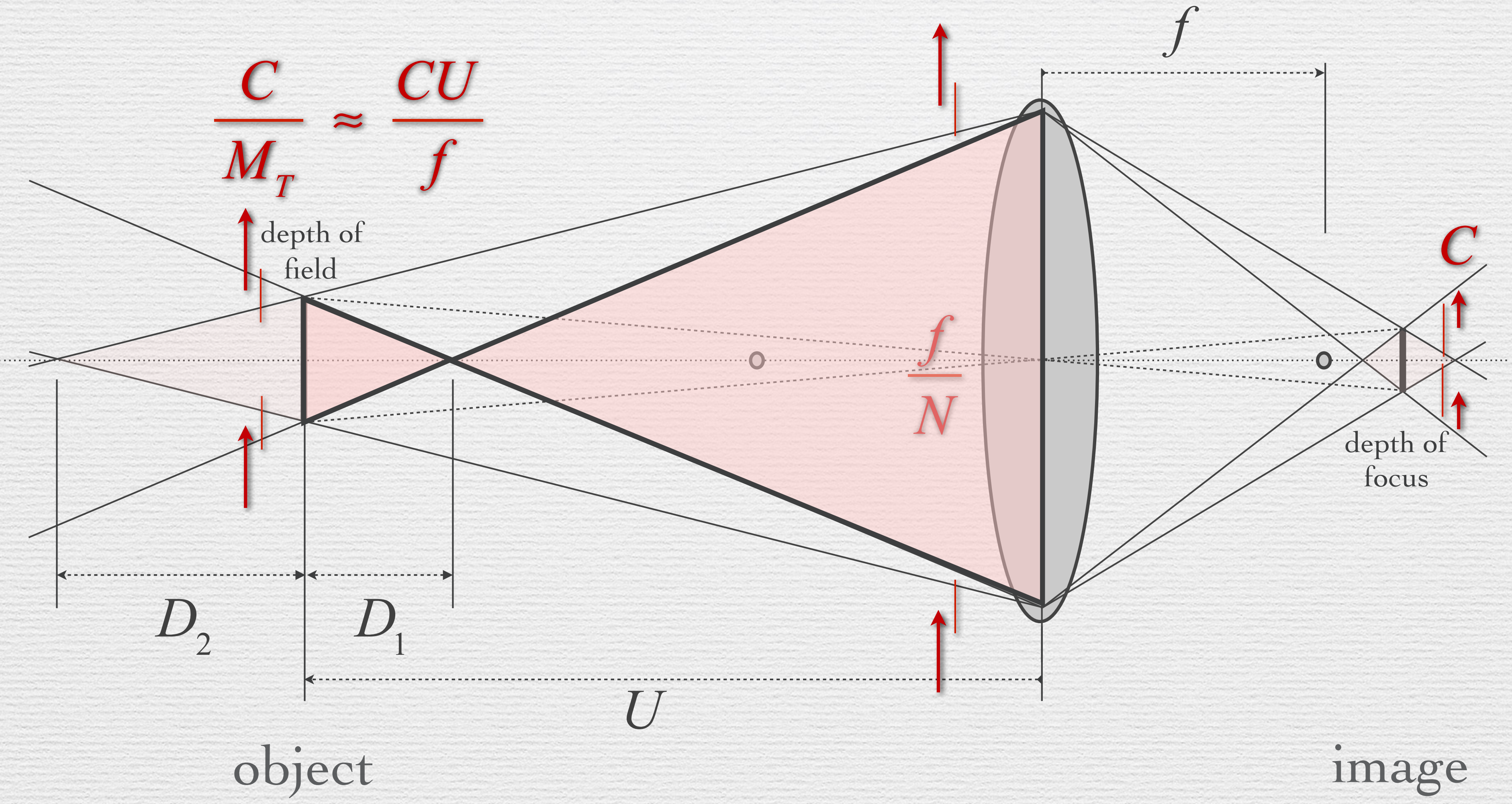


DXOMARK
MOBILE

98	Google Pixel 2
97	Apple iPhone X
97	Huawei Mate 10 Pro
94	Apple iPhone 8 Plus
94	Samsung Galaxy Note 8
92	Apple iPhone 8
90	Google Pixel
90	HTC U11
90	Xiaomi Mi Note 3
88	Apple iPhone 7 Plus
85	Apple iPhone 7
83	Sony Xperia XZ Premium

Depth of field formula

[https://sites.google.com/site/marclevoylectures/]



$$\frac{D_1}{CU/f} = \frac{U - D_1}{f/N} \quad \ldots\ldots \quad D_1 = \frac{NCU^2}{f^2 + NCU} \qquad D_2 = \frac{NCU^2}{f^2 - NCU}$$

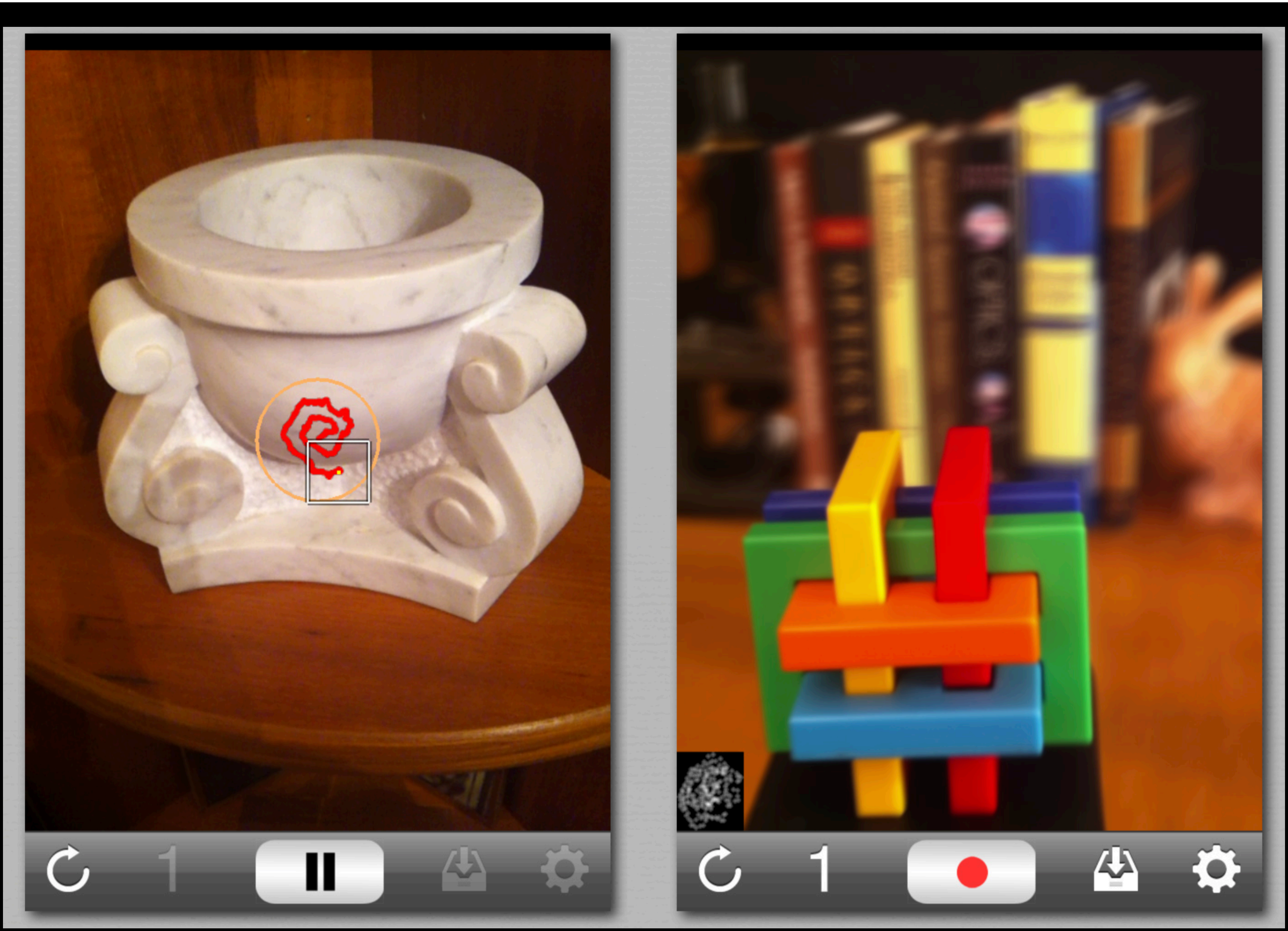
Depth of field formula

$$\boxed{D_{TOT} \approx \frac{2NCU^2}{f^2}}$$

- where
  - N is F-number of lens
  - C is circle of confusion (on image)
  - U is distance to in-focus plane (in object space)
  - f is focal length of lens

$$D_{TOT} \approx \frac{2NCU^2}{f^2}$$

- N = f/4.1
  C = 2.5µ
  U = 5.9m (19’)
  f = 73mm (equiv to 362mm)
  D_{TOT} = 132mm

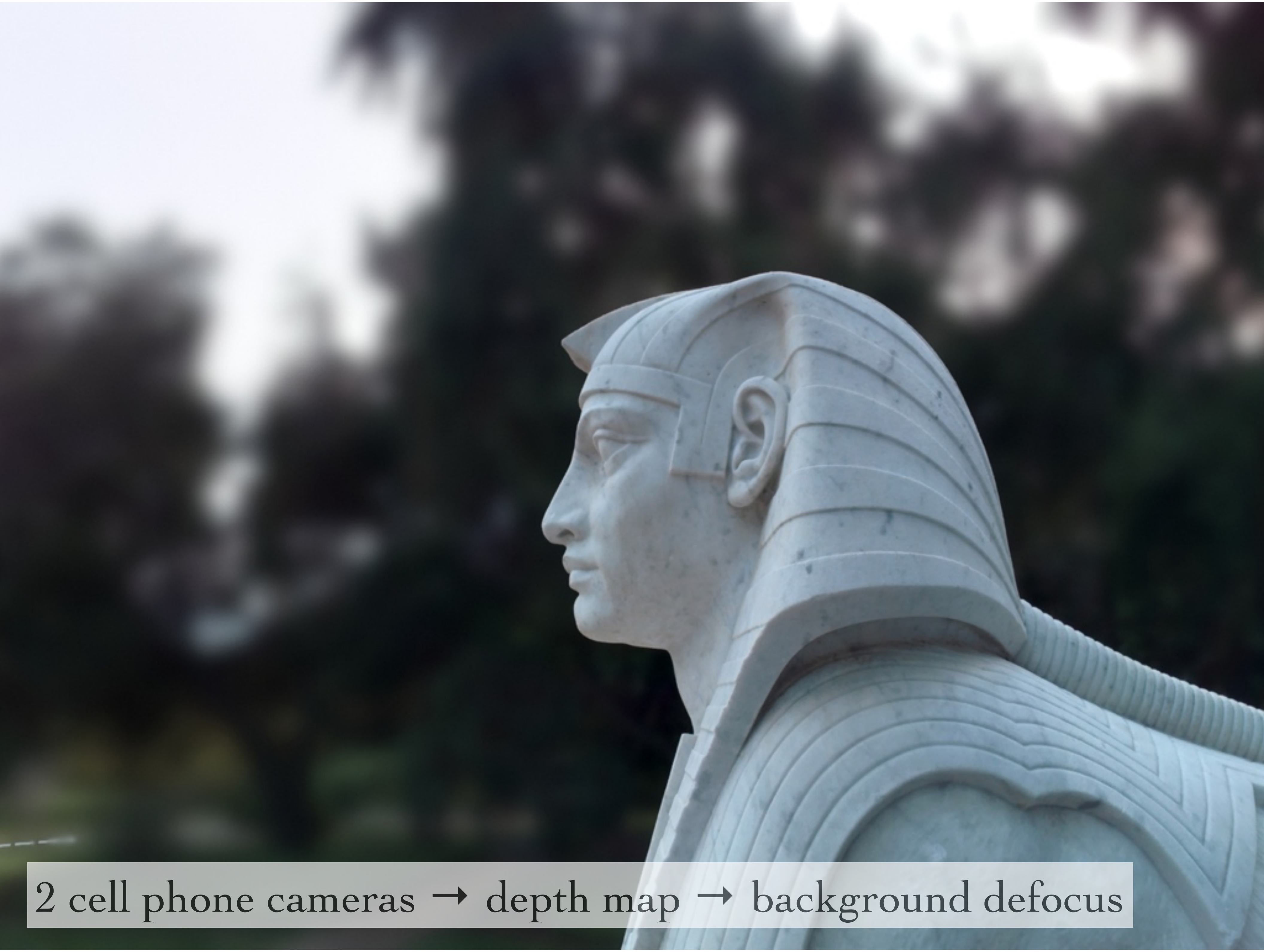


SynthCam: discretely approximated real depth of field

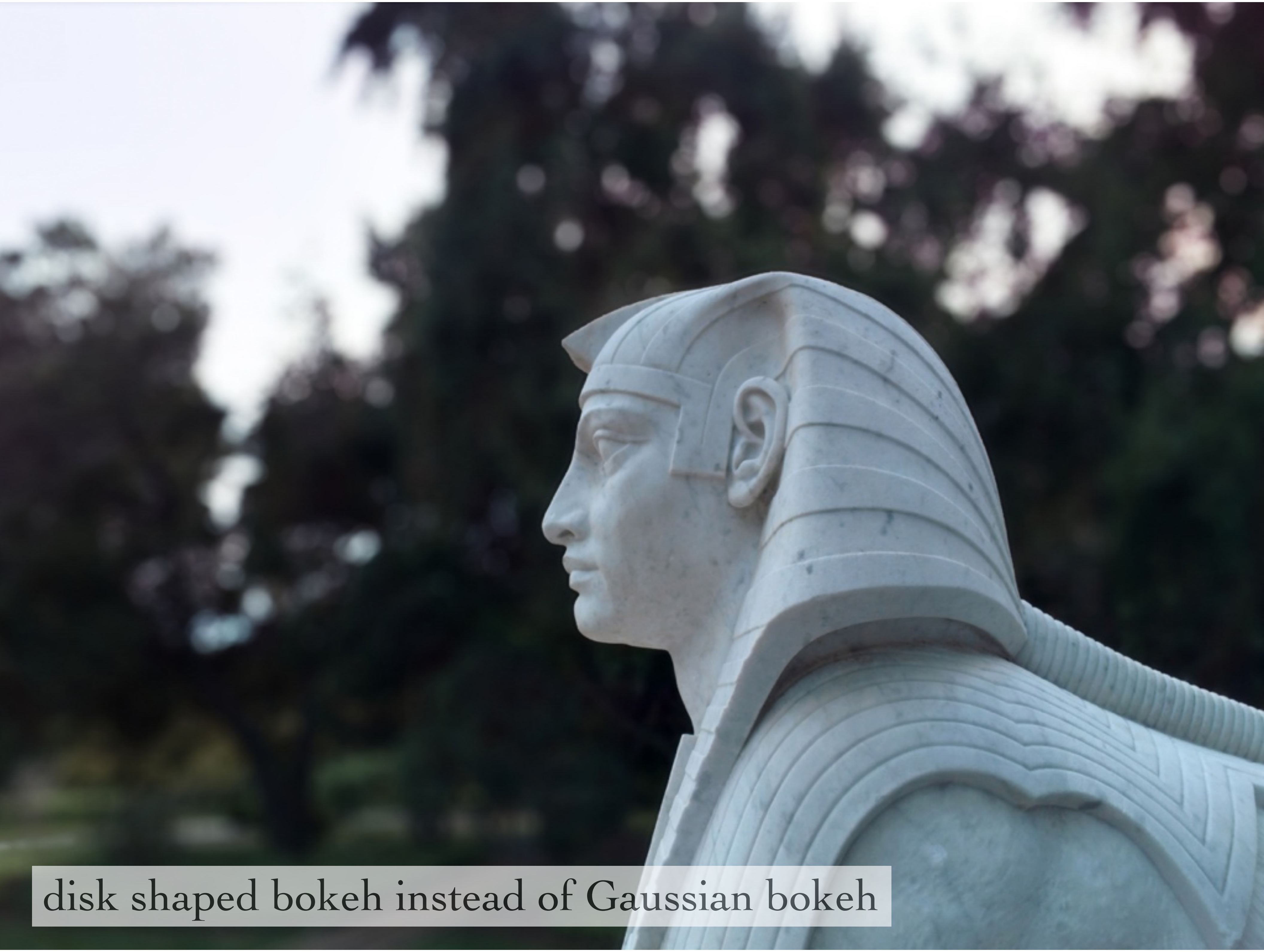
Synthetic shallow depth of field

- dual-camera phones
- capture two images with similar viewpoints
- use stereo matching to compute a depth map
- choose one plane in the scene to keep sharp
- blur features that are closer or further away

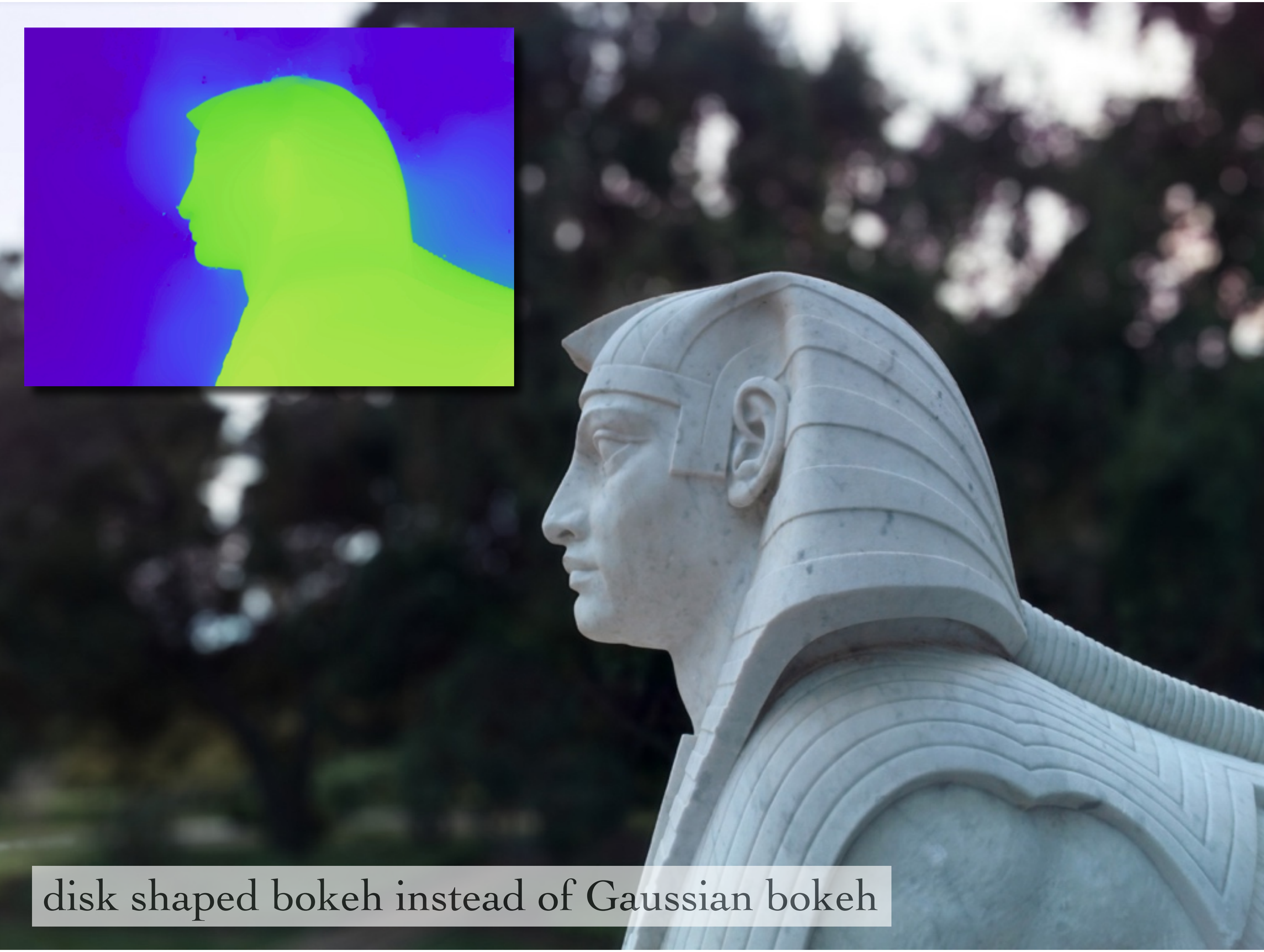
cell phone camera



2 cell phone cameras → depth map → background defocus



disk shaped bokeh instead of Gaussian bokeh



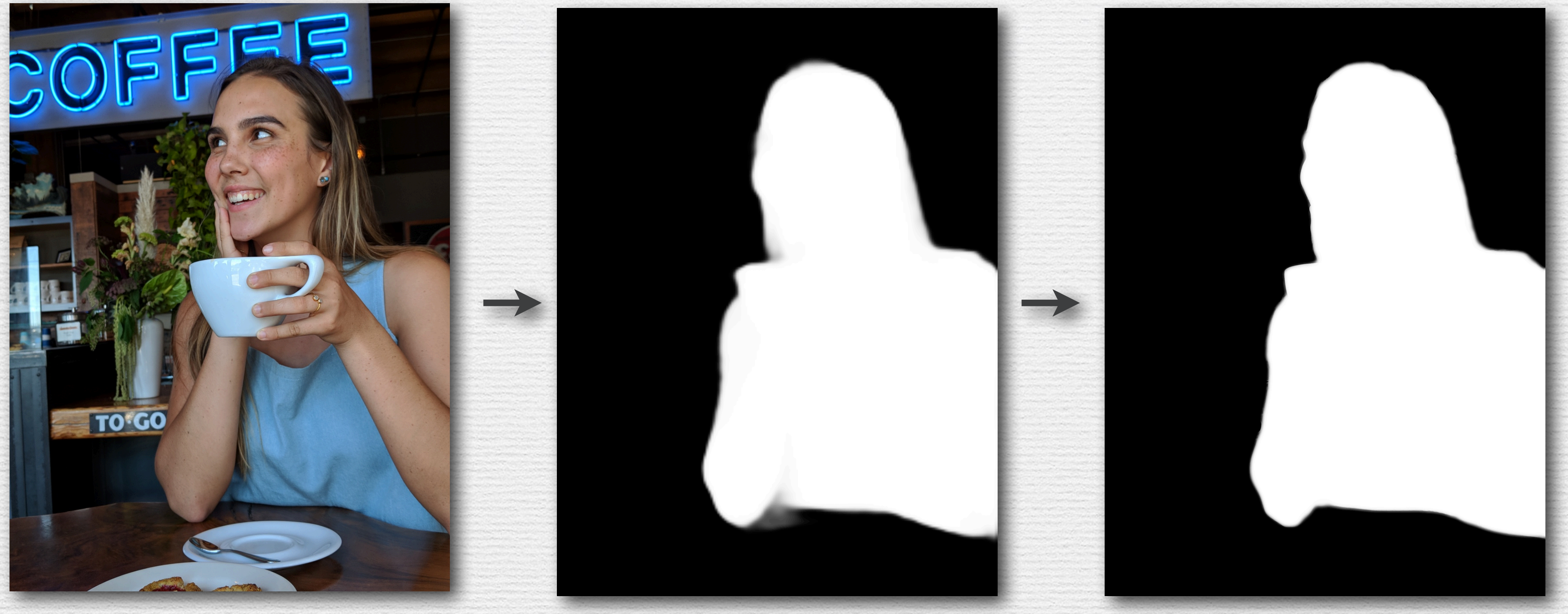
disk shaped bokeh instead of Gaussian bokeh

But the Pixel 2 has only one camera!

1. use machine learning to segment people
2. use dual-pixels to estimate a depth map

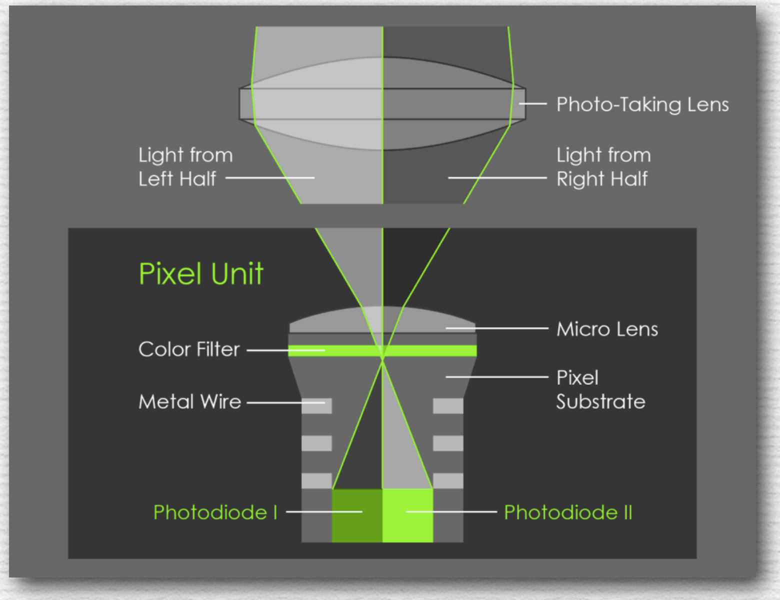
- combine these two signals
- for selfie camera use only #1
- for macro objects use only #2

1. Learning-based segmentation



- ✦ CNN estimates prob(person) at every pixel
  - trained on 1M labeled pictures of people and accessories
  - synthetic training data (one person, multiple backgrounds)
- ✦ refined using edge-aware bilateral solver
  [Barron and Poole, ECCV 2016]

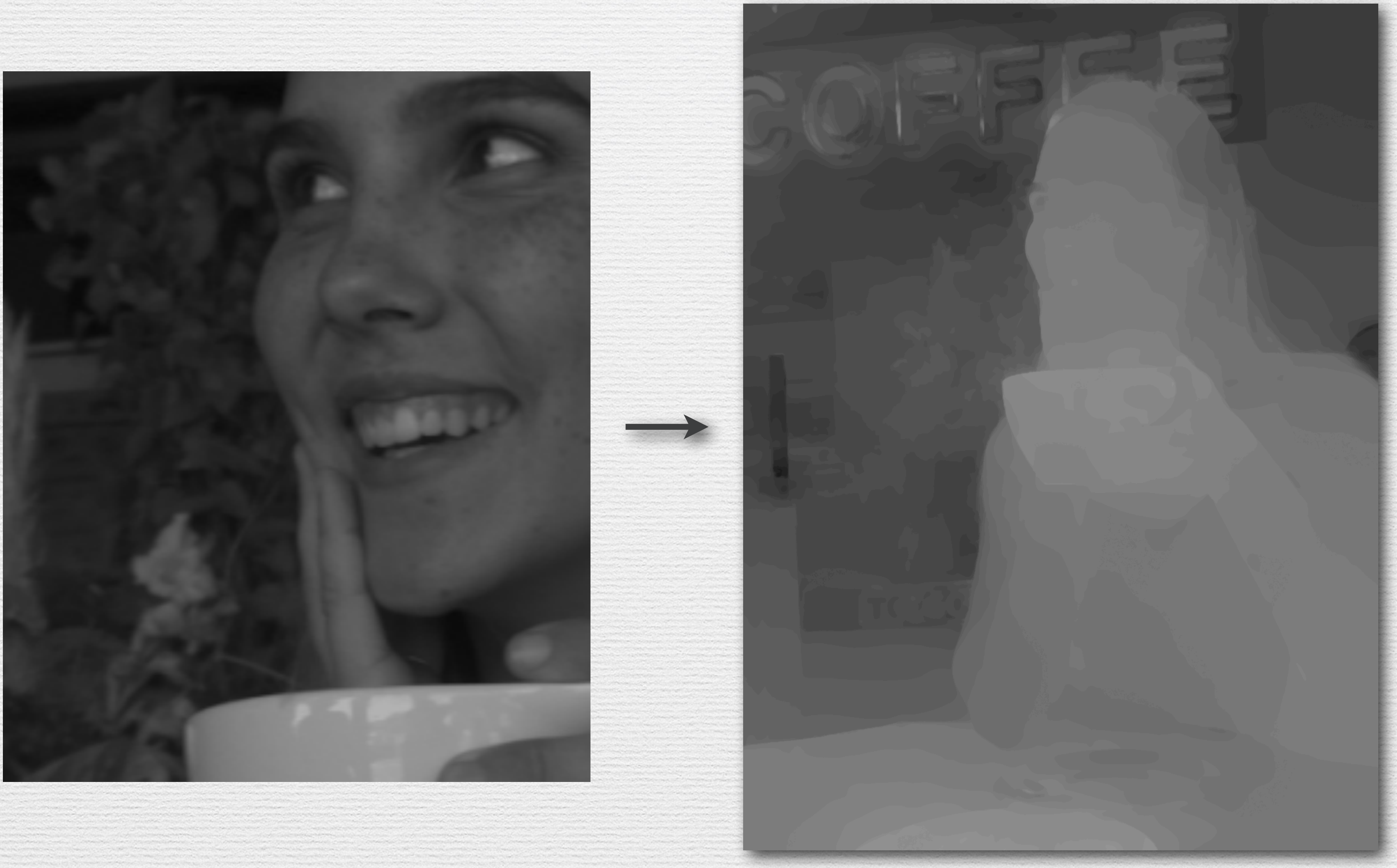
2. Dual pixels



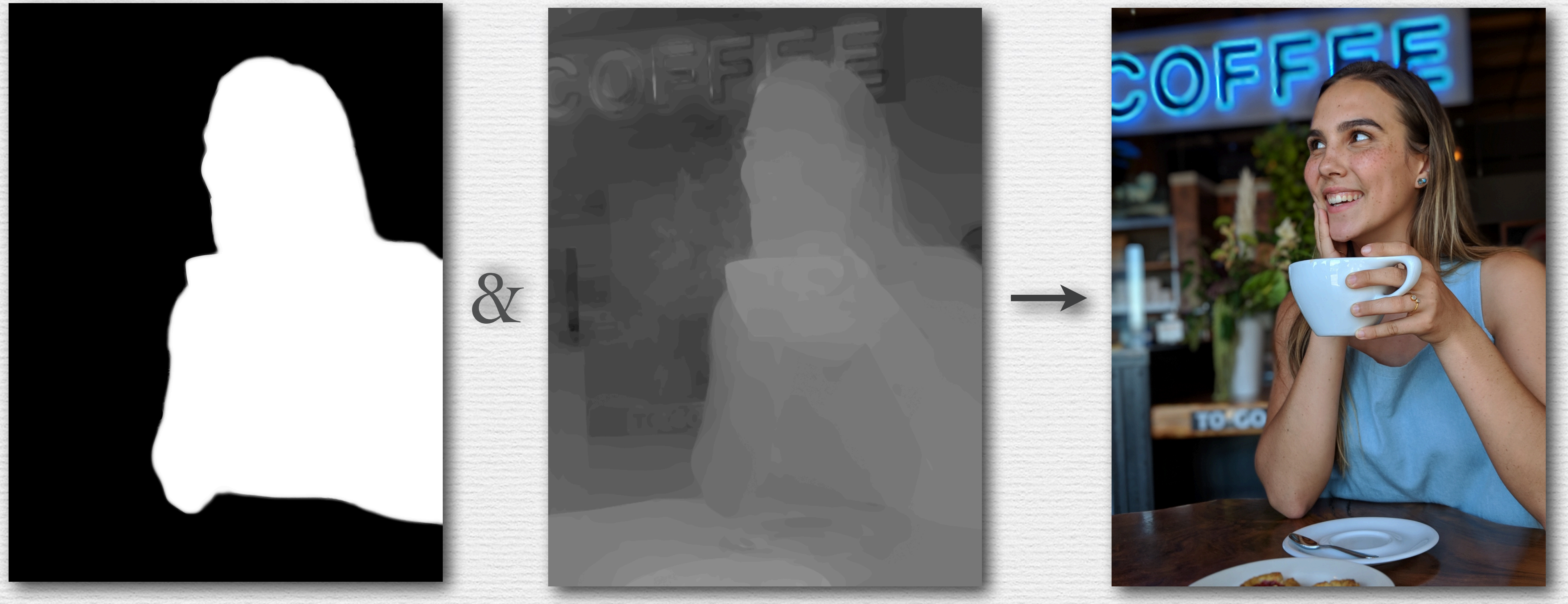
(Markus Kohlpaintner)

- a.k.a. phase-detect auto-focus (PDAF)
- used to focus while video recording in newer SLRs
- each pixel is split in half
- left half sees through right half of lens
- stereo with a very tiny baseline (1mm)

Depth from dual pixels

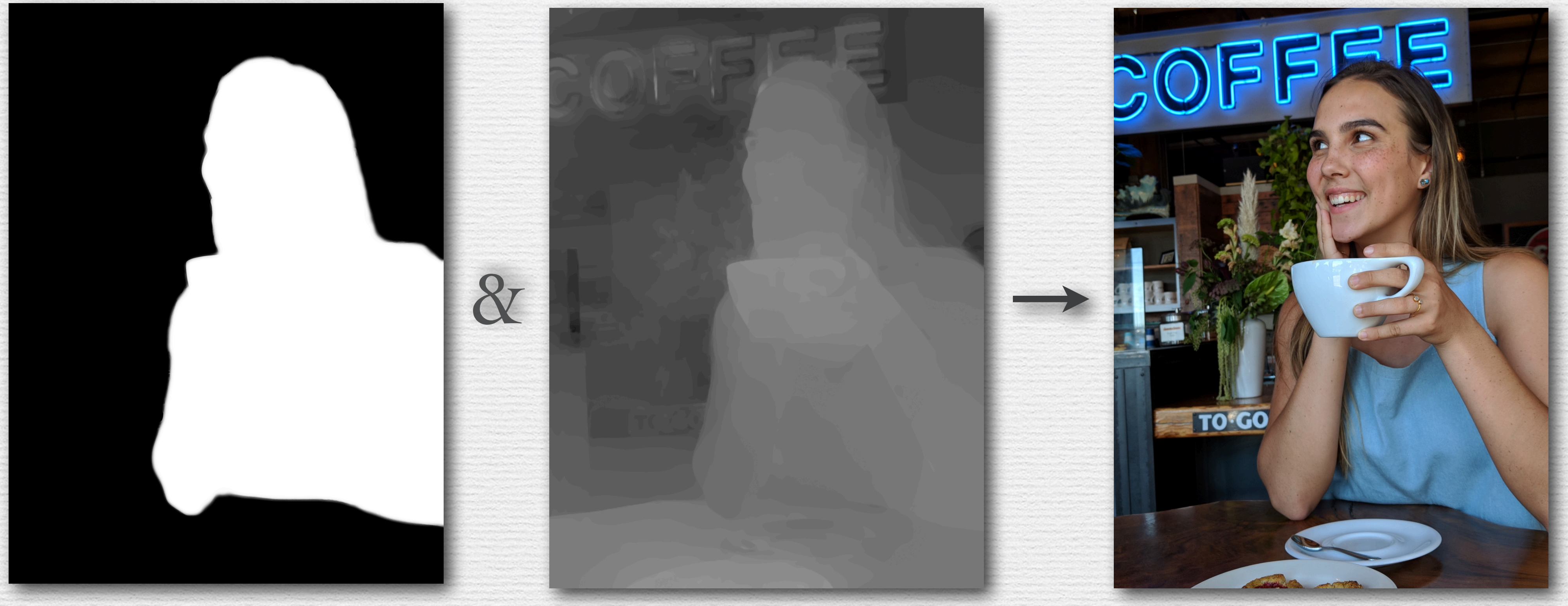


Blurring based on mask and depth



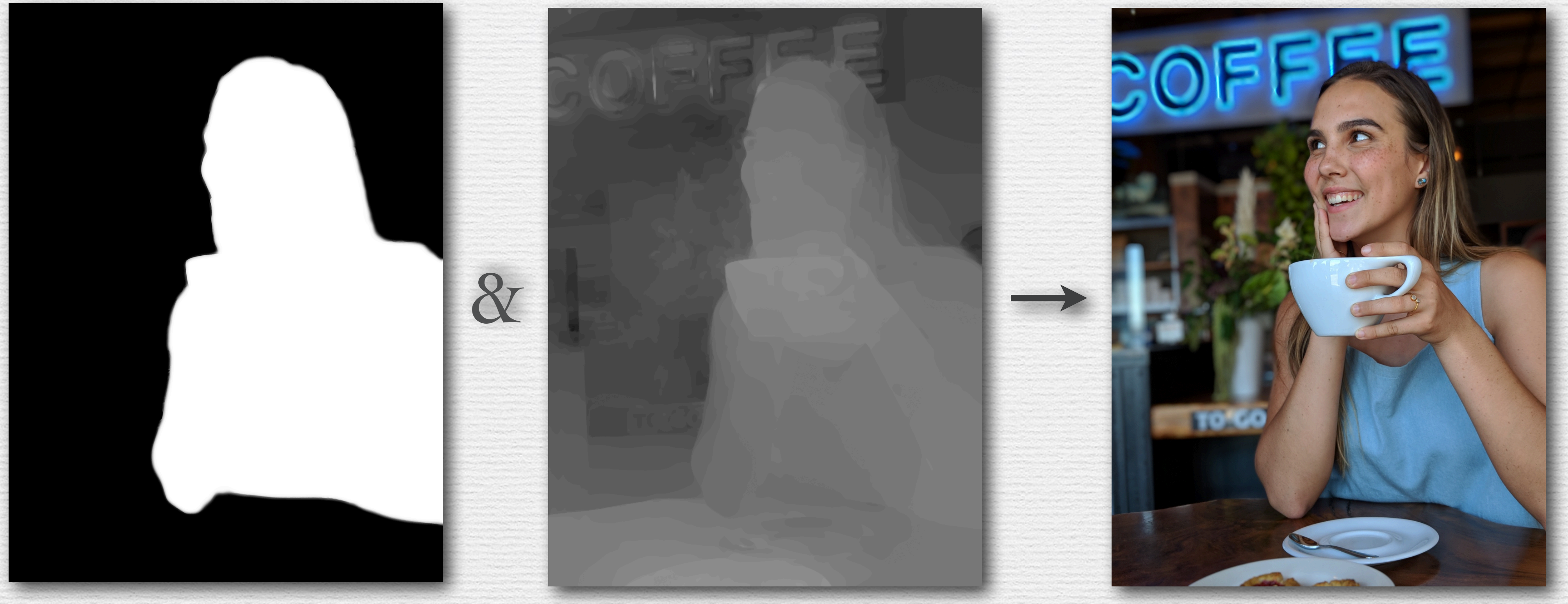
- ✦ keep entire person sharp
- ✦ blur proportional to distance from person
- ✦ keep a zone of depths around person sharp
  - not physically correct, but helps novices take portraits

Blurring based on mask and depth



- keep entire person sharp
- blur proportional to distance from person
- keep a zone of depths around person sharp
  - not physically correct, but helps novices take portraits

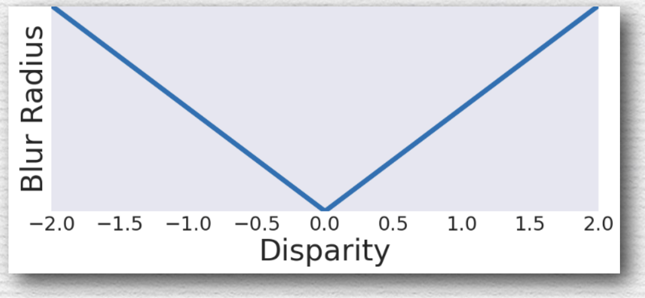
Blurring based on mask and depth



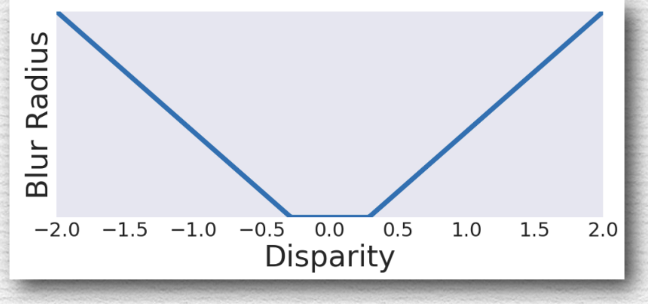
- keep entire person sharp
- blur proportional to distance from person
- keep a zone of depths around person sharp
  - not physically correct, but helps novices take portraits

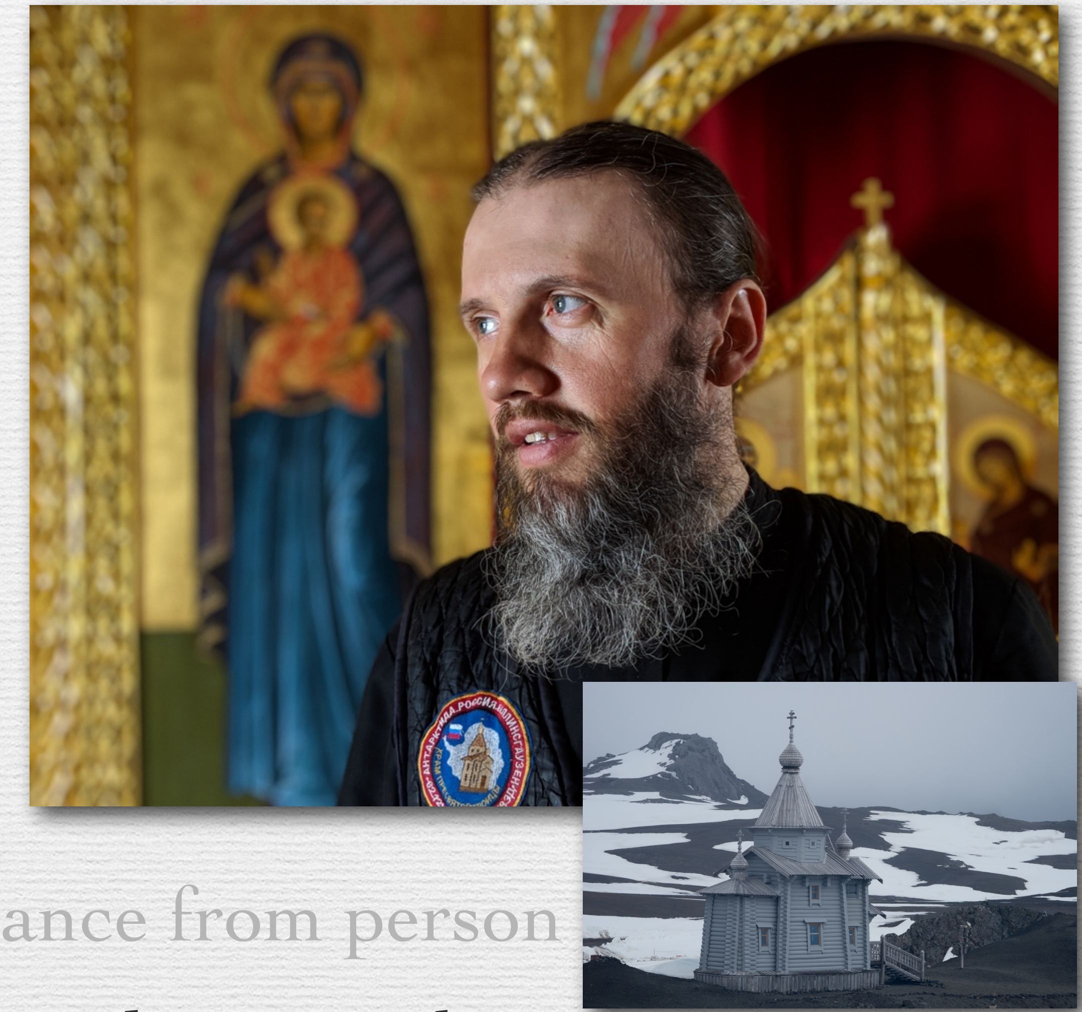
Blurring based on mask and depth

correct



extended



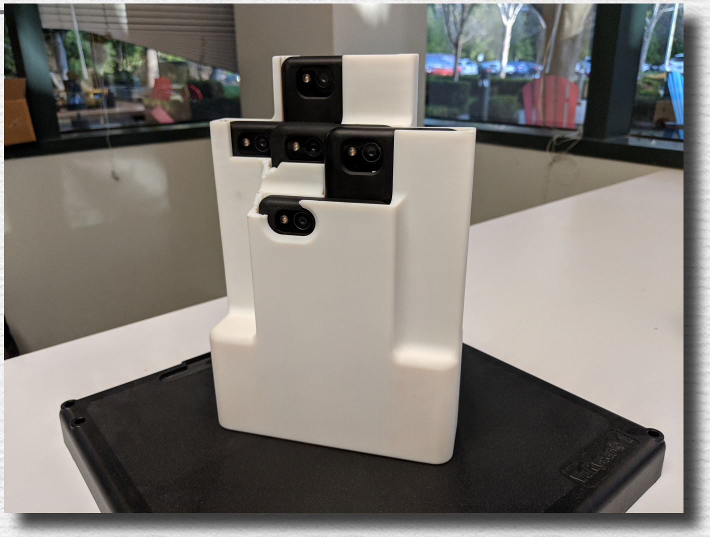


- ✦ keep entire person sharp
- ✦ blur in proportion to distance from person
- ✦ keep a zone of depths around person sharp
  - not physically correct, but helps novices take portraits

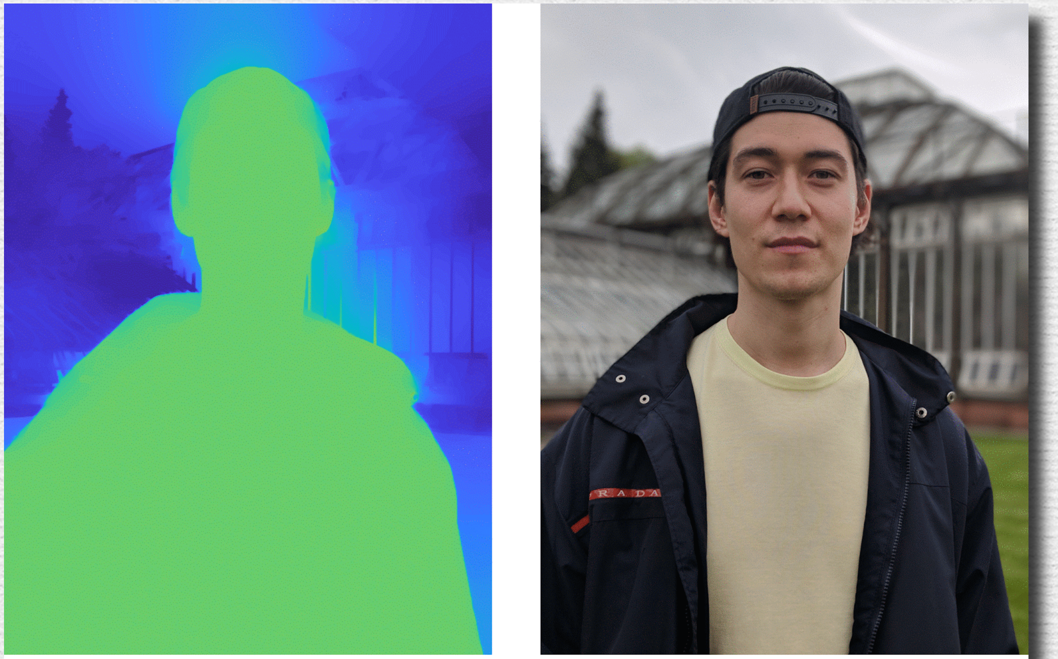
New in Pixel 3: learning-based depth-from-dual pixels

[Garg and Wadhwa, Google AI blog]

- input is R, G, B, left, right
- output is depth map
- ground truth is better depth map from a multi-camera stereo rig ⟶

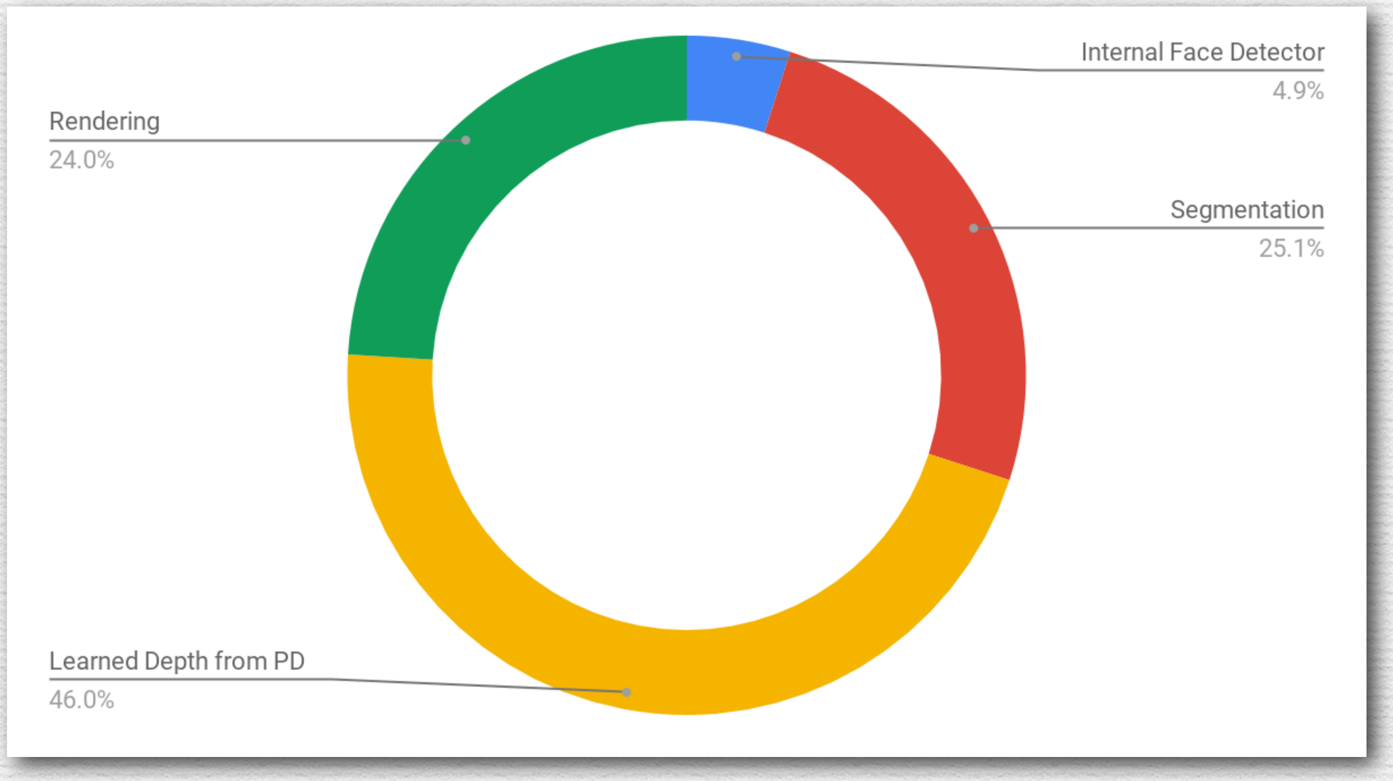


Original



Stereo Depth

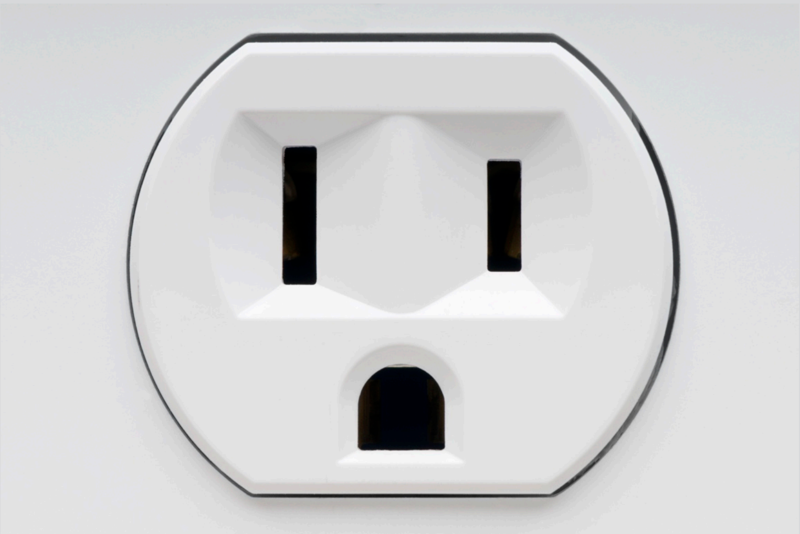
Performance



- HDR+ (2 secs) + portrait (2 secs) = 4 seconds
- 50x too slow for live bokeh effect in the viewfinder

Where else can ML be used?

- feasible and likely
  - face detection, object recognition, scene recognition



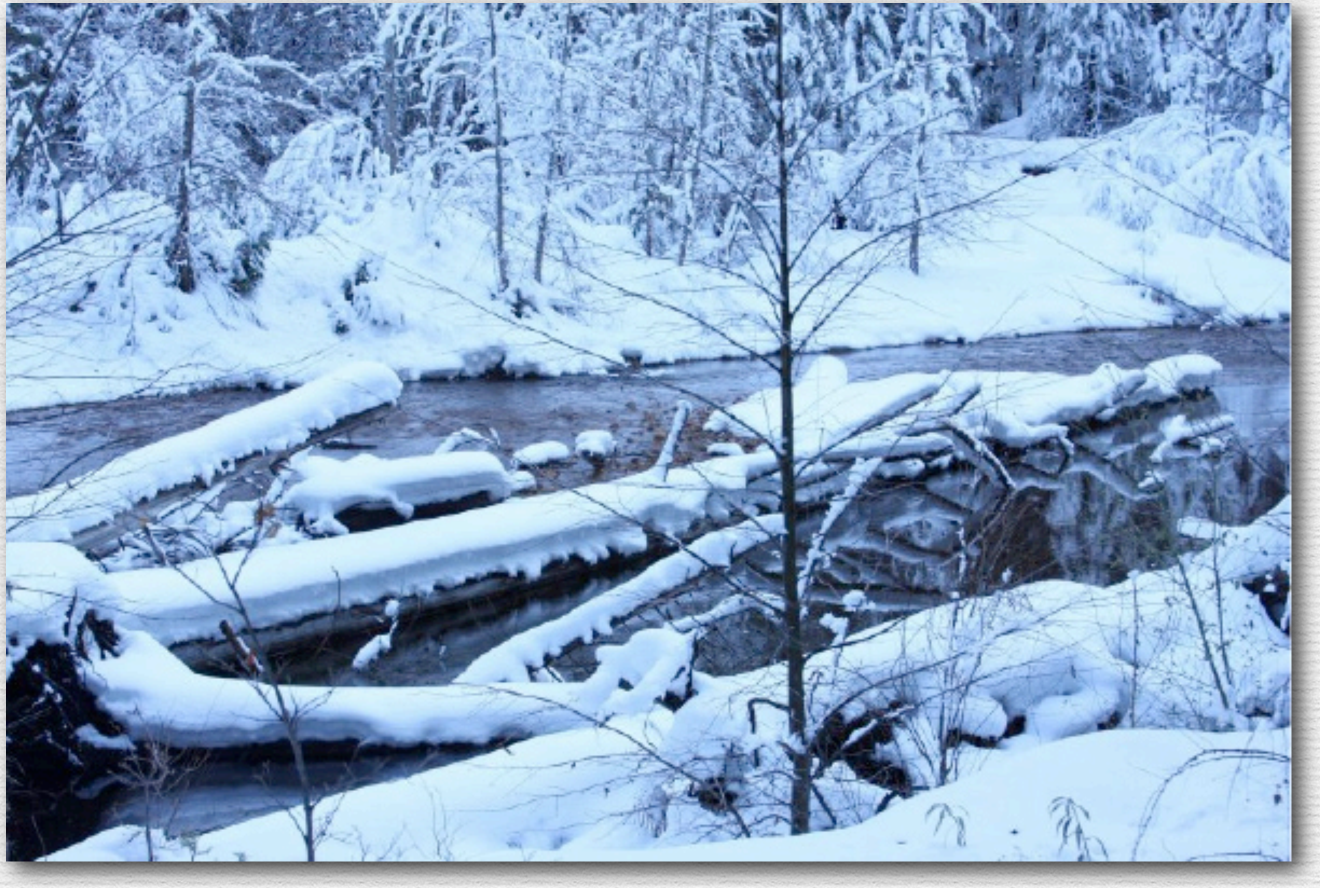
Ceci n’est pas une face

Ceci n'est pas une pipe.

Where else can ML be used?

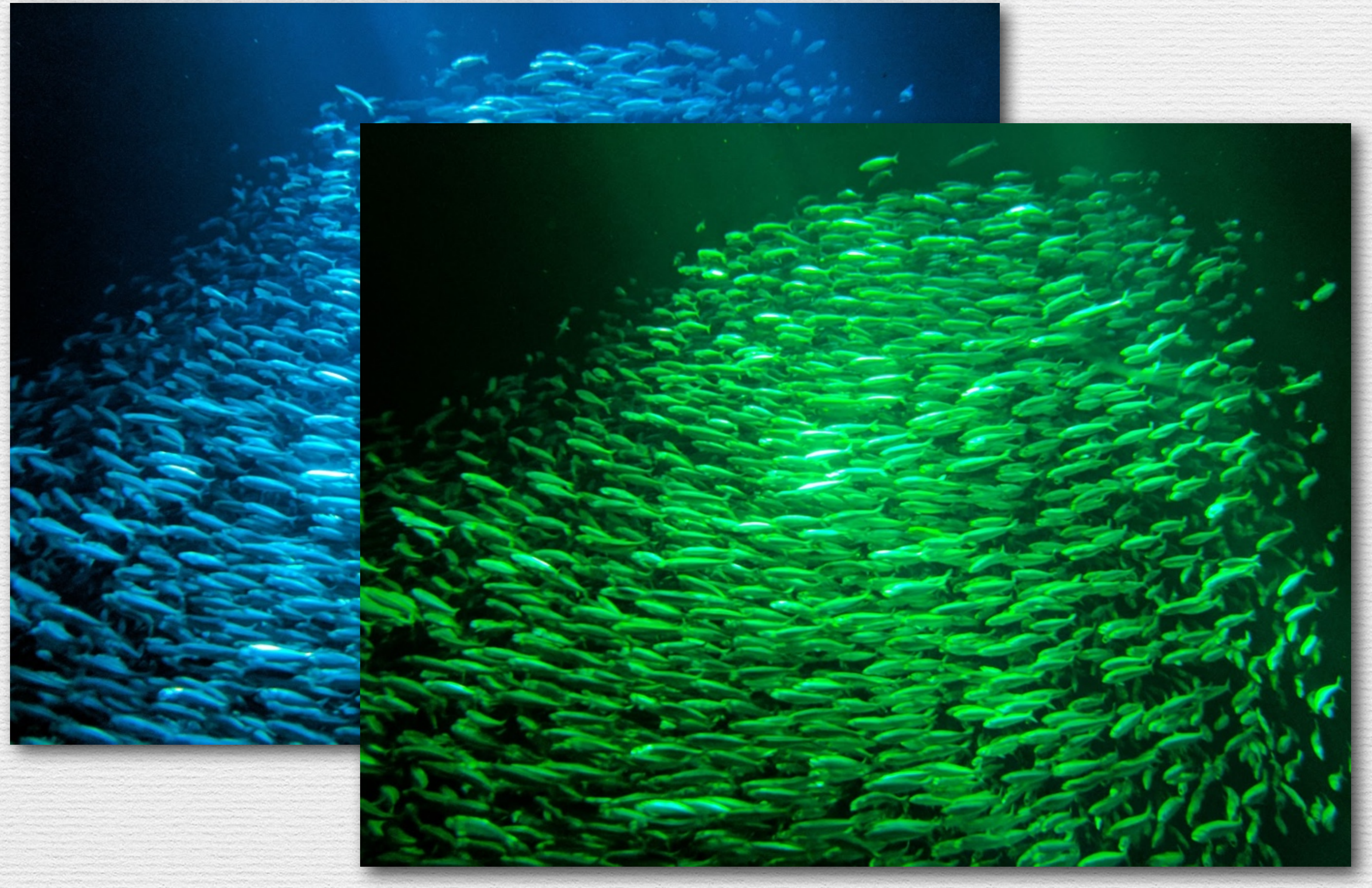
- feasible and likely
  - face detection, object recognition, scene recognition
  - 3A (auto-exposure, auto-focus, auto-white-balance)

White balancing is an ill-posed problem



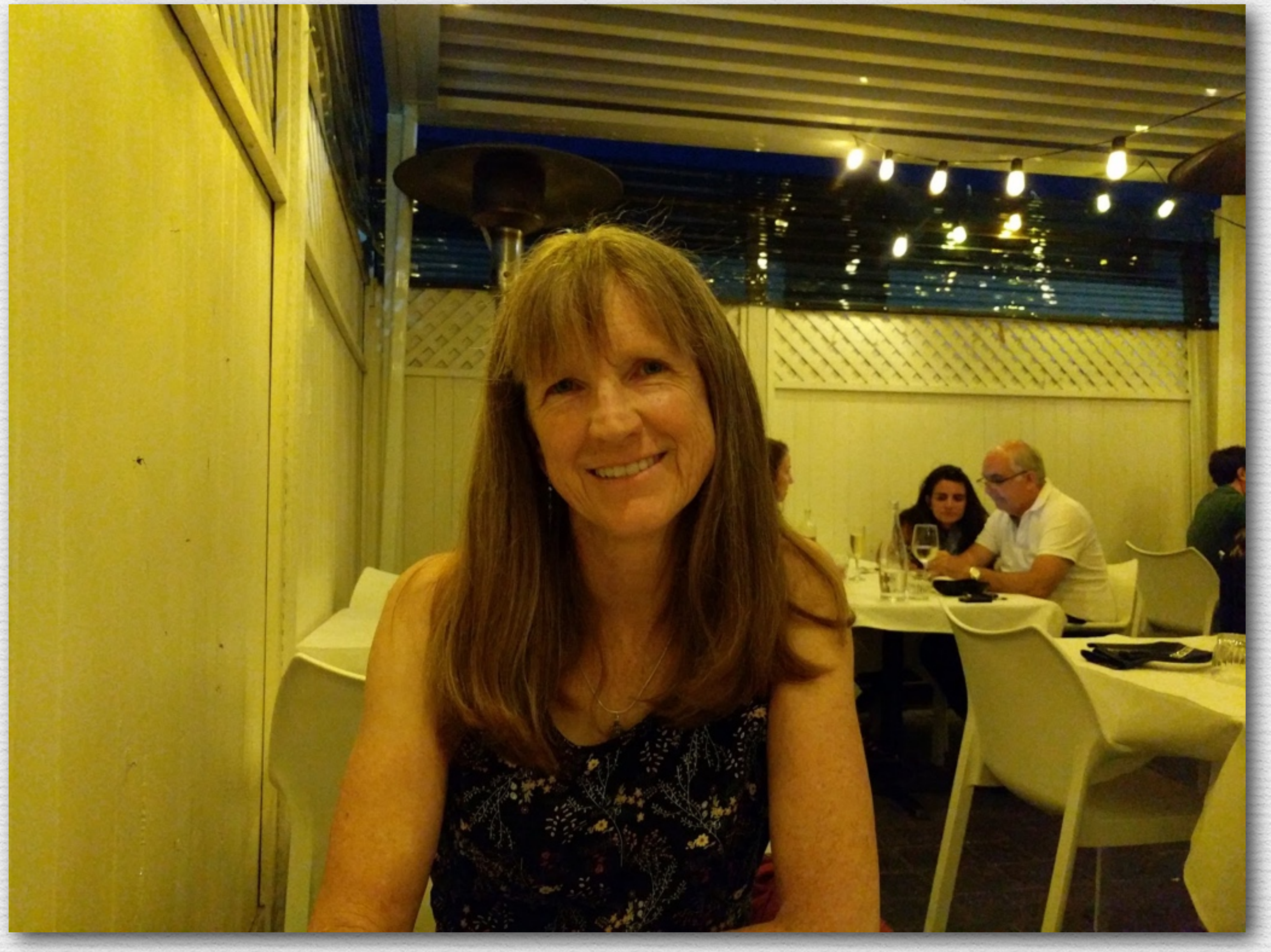
- Is this blue snow?
- Or white snow illuminated by a blue sky?

Typical white balancing failures



- green is not a likely color for fish in an aquarium tank

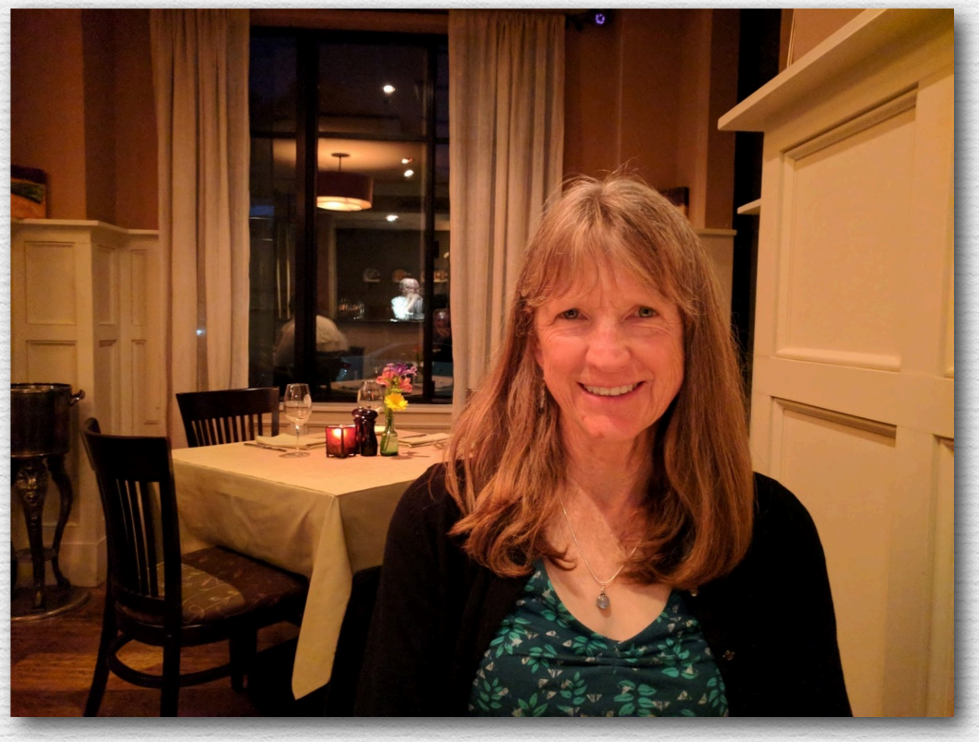
Typical white balancing failures

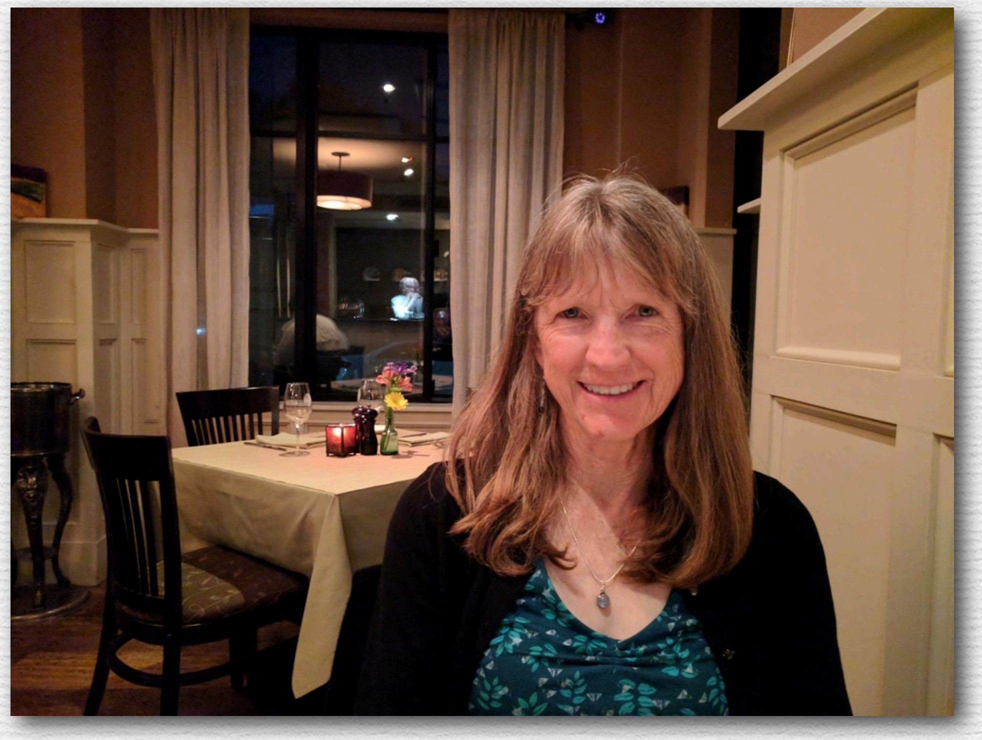


- yellow is not a likely color for human skin

Learning-based white balancing

[Barron ICCV 2015, Barron CVPR 2017]

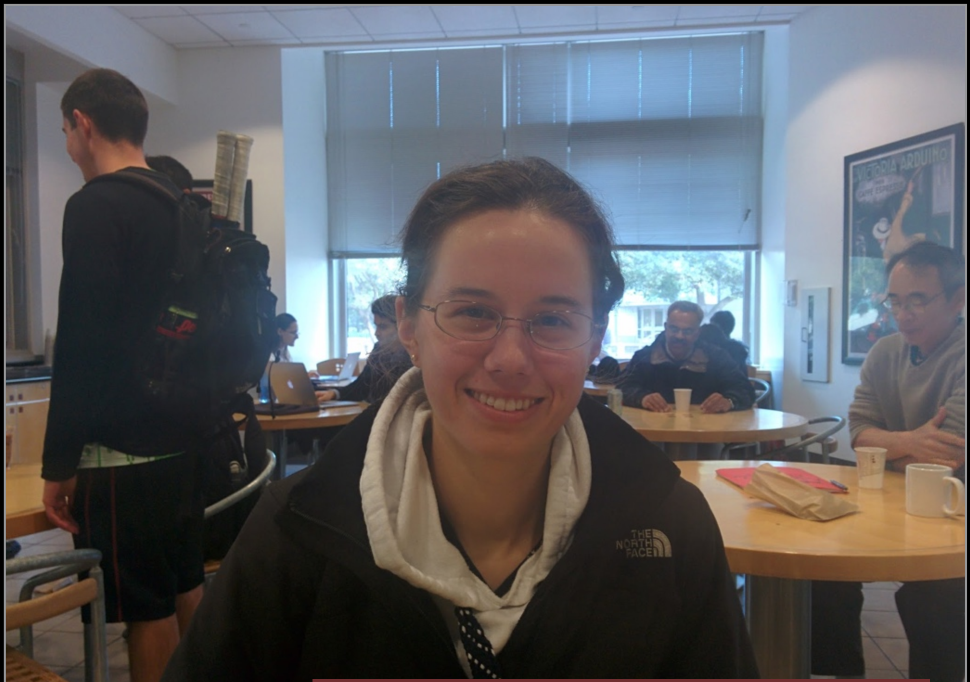


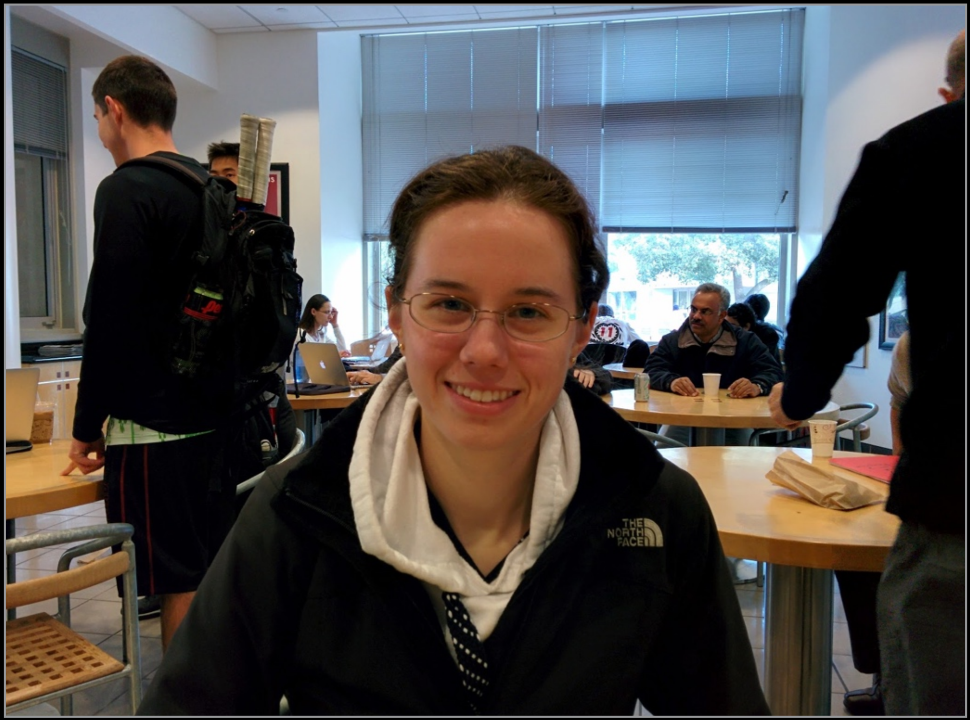


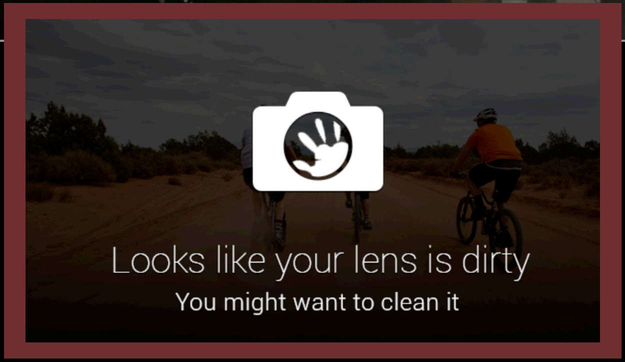
- training data is well-balanced images
  - manually tagged or scraped from existing collections

Where else can ML be used?

- feasible and likely
  - face detection, object recognition, scene recognition
  - 3A (auto-exposure, auto-focus, auto-white-balance)
- hard or impossible
  - curation
  - super-resolution
  - photographer's assistant

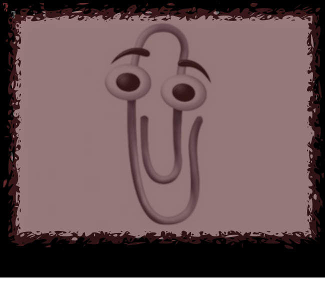






Looks like your lens is dirty

You might want to clean it



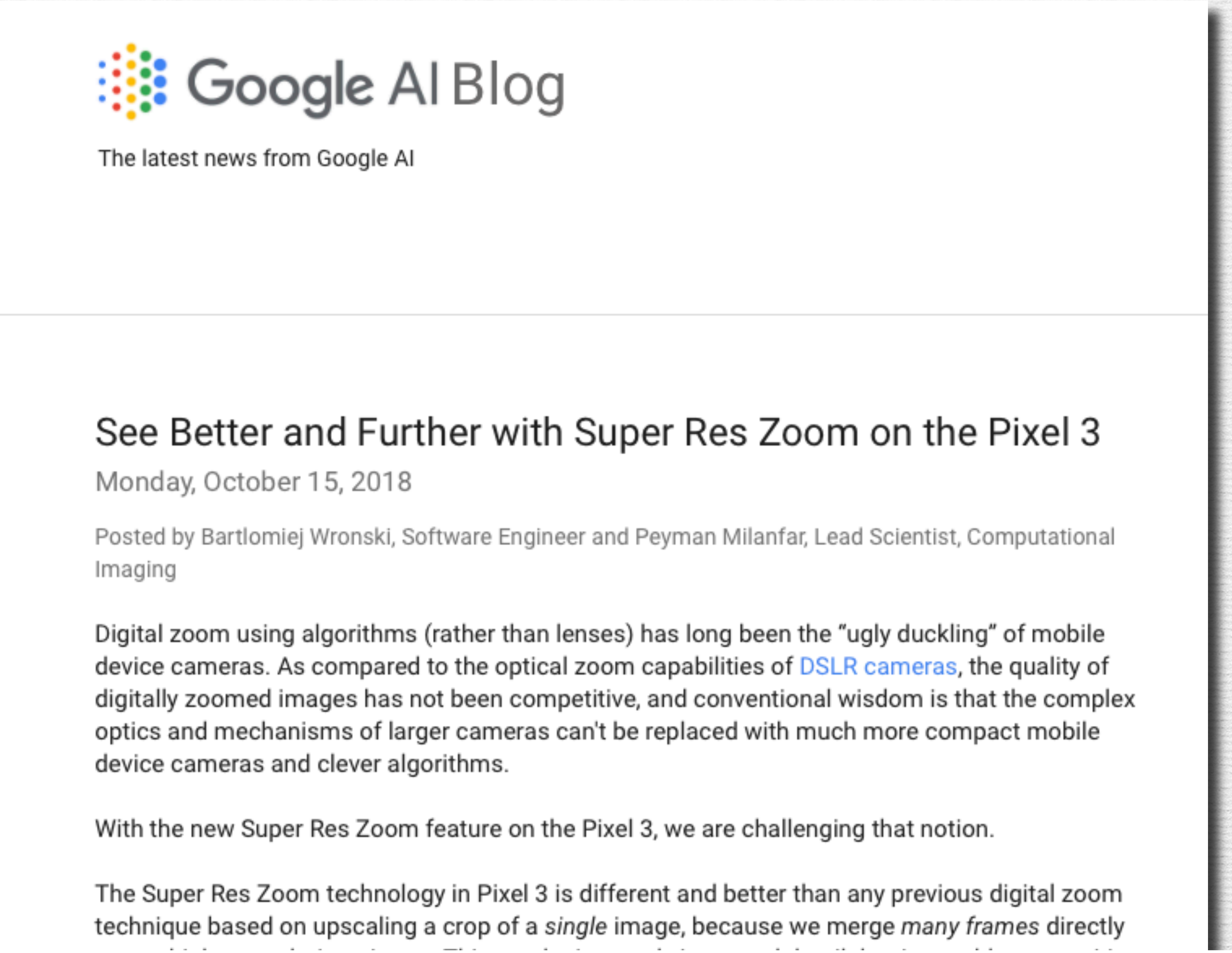
How far can cell phone cameras go?

- ✦ ways in which SLRs beat cell phones
  - ✓ • dynamic range (in bright scenes)
  - ✓ • signal-to-noise (in dark scenes)
  - ✓ • shallow depth of field
  - ?✗? • narrow field of view (i.e. telephoto)

? ?

Example #3: Super Res Zoom on Pixel 3

[Wronski and Milanfar, Google AI blog]



Google AI Blog

The latest news from Google AI

See Better and Further with Super Res Zoom on the Pixel 3

Monday, October 15, 2018

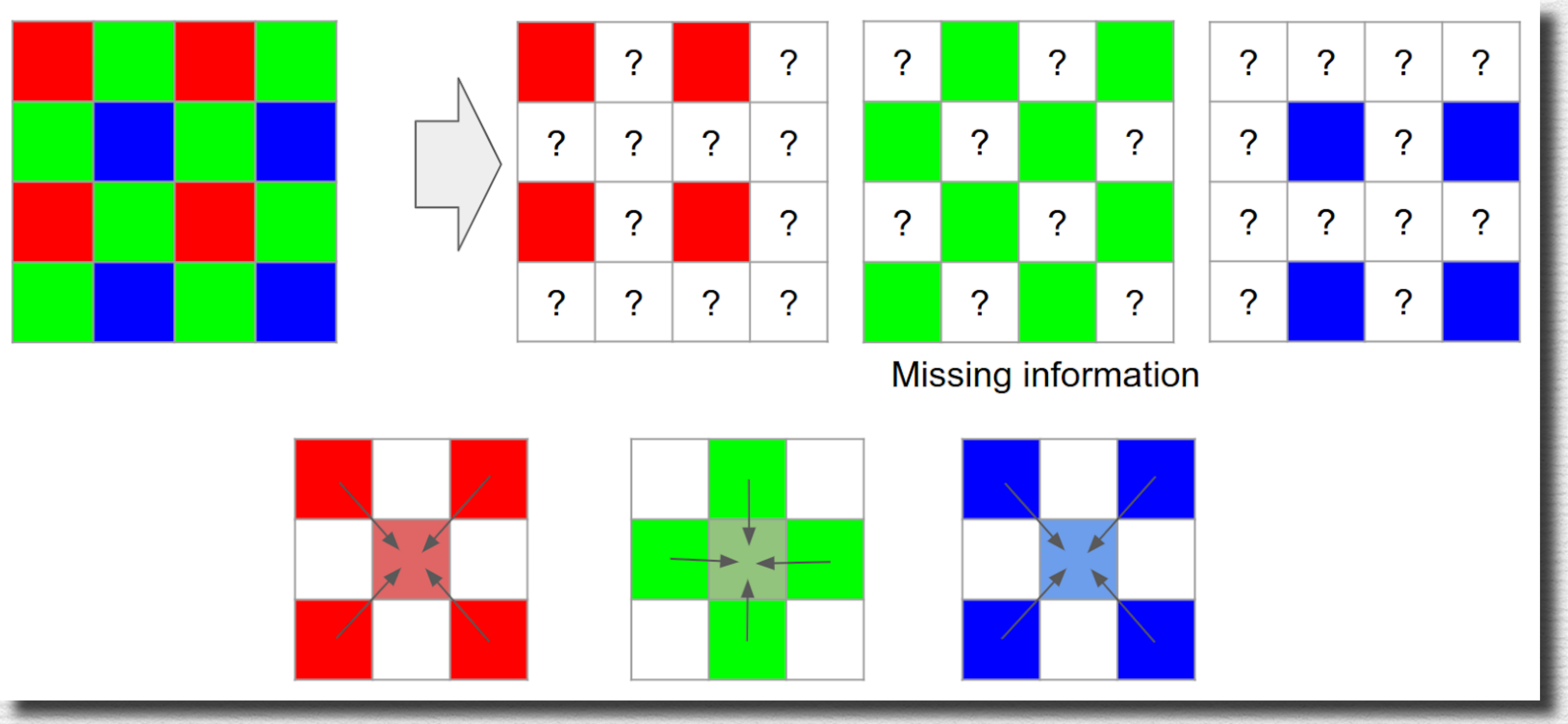
Posted by Bartlomiej Wronski, Software Engineer and Peyman Milanfar, Lead Scientist, Computational Imaging

Digital zoom using algorithms (rather than lenses) has long been the "ugly duckling" of mobile device cameras. As compared to the optical zoom capabilities of DSLR cameras, the quality of digitally zoomed images has not been competitive, and conventional wisdom is that the complex optics and mechanisms of larger cameras can't be replaced with much more compact mobile device cameras and clever algorithms.

With the new Super Res Zoom feature on the Pixel 3, we are challenging that notion.

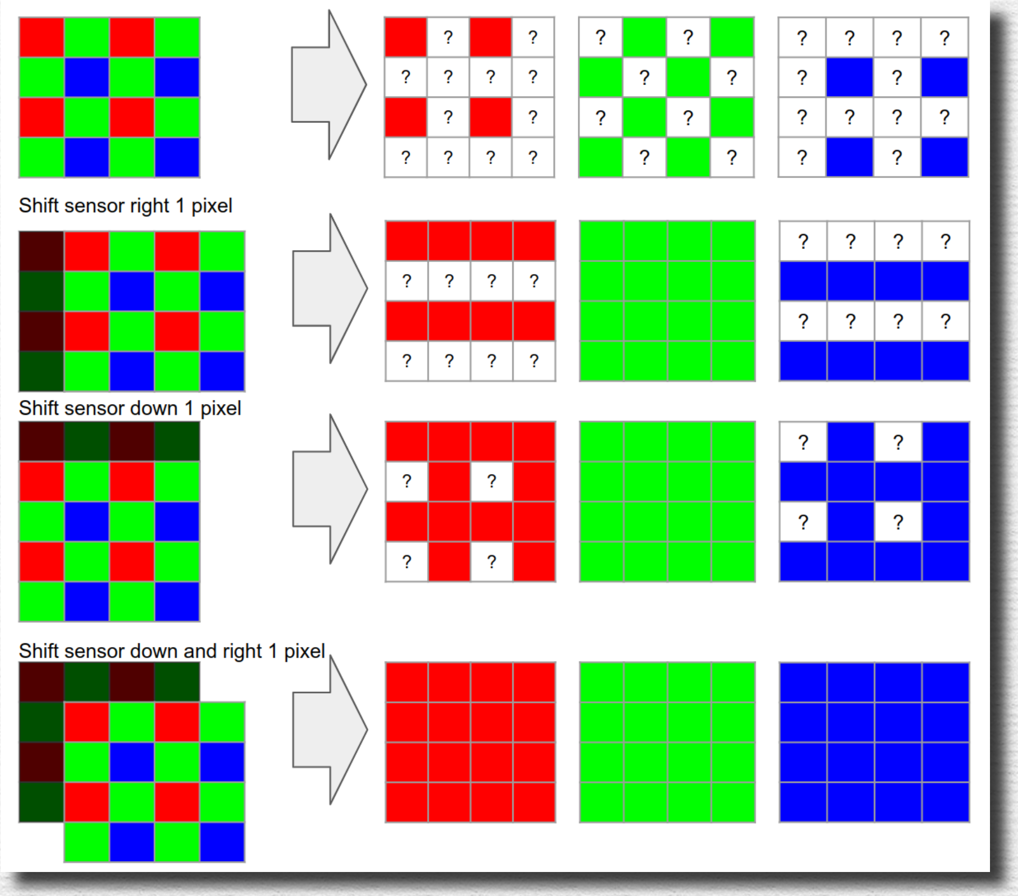
The Super Res Zoom technology in Pixel 3 is different and better than any previous digital zoom technique based on upscaling a crop of a *single* image, because we merge *many frames* directly

Demosaicing versus pixel shifting



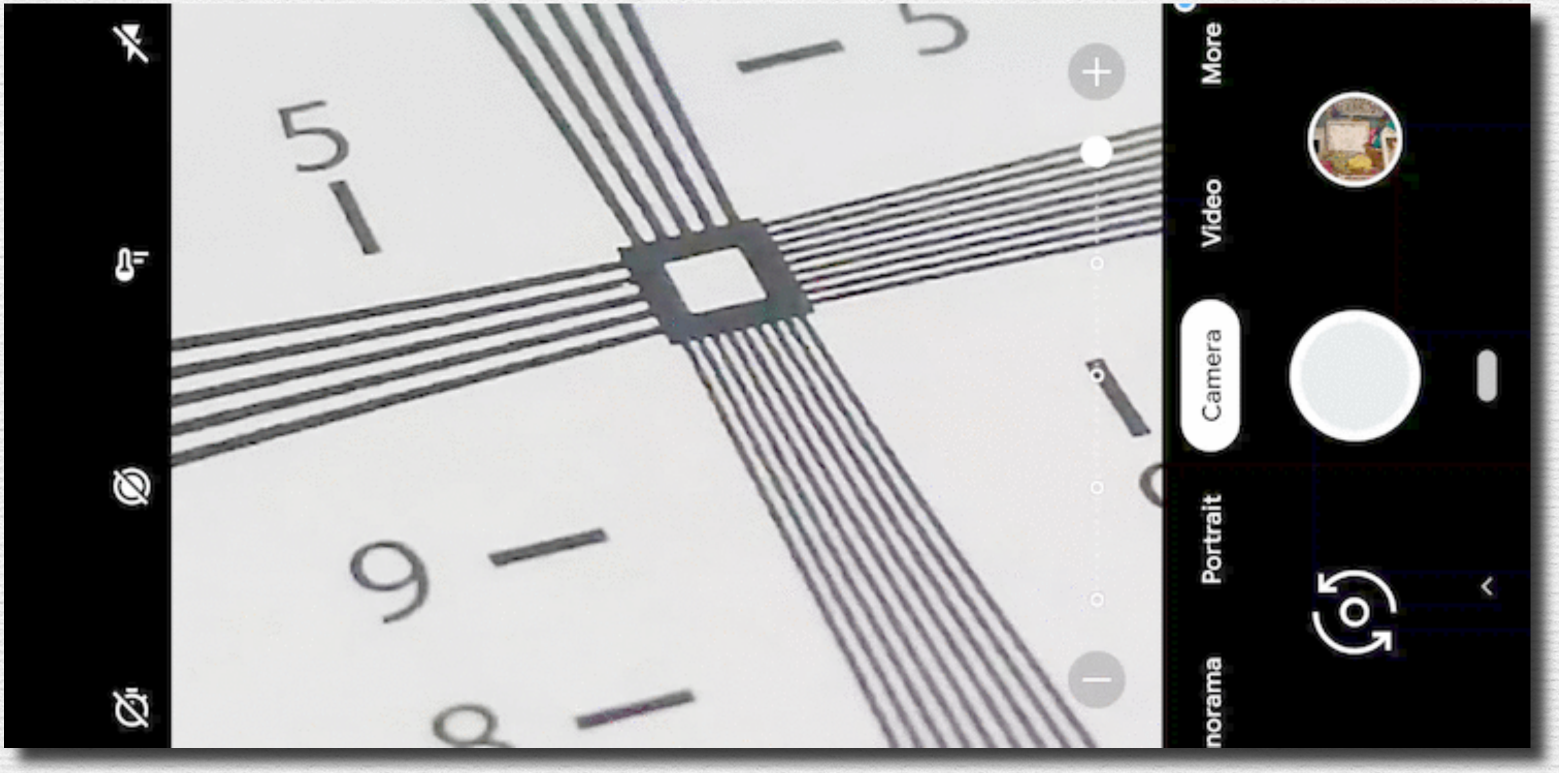
- must interpolate red, green, and/or blue at most pixels
- 2/3 of your picture is made up!

Demosaicing versus pixel shifting


Shift sensor right 1 pixel
Shift sensor down 1 pixel
Shift sensor down and right 1 pixel
?

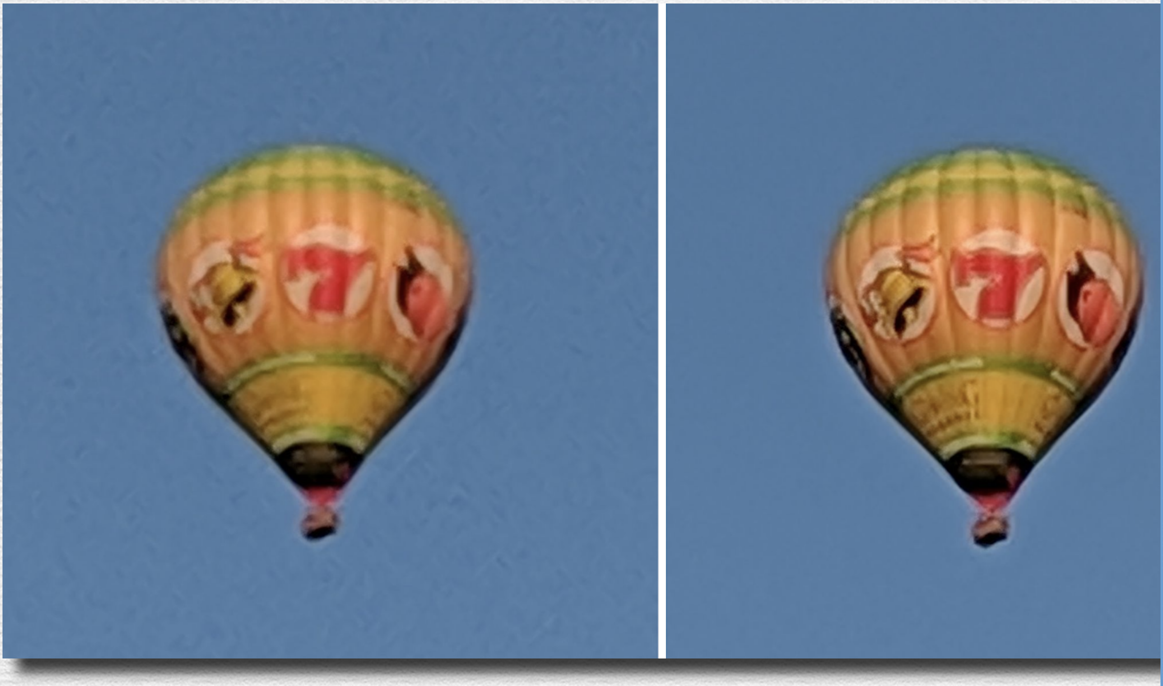
- SLRs on tripods use pixel shifting to avoid demosaicing
- if handheld, use handshake and alignment instead!

What if your hands are “too steady”?



- wiggle the optical image stabilizer (OIS) between frames
  [Ben-Ezra, Zomet, Nayar, CVPR 2004]

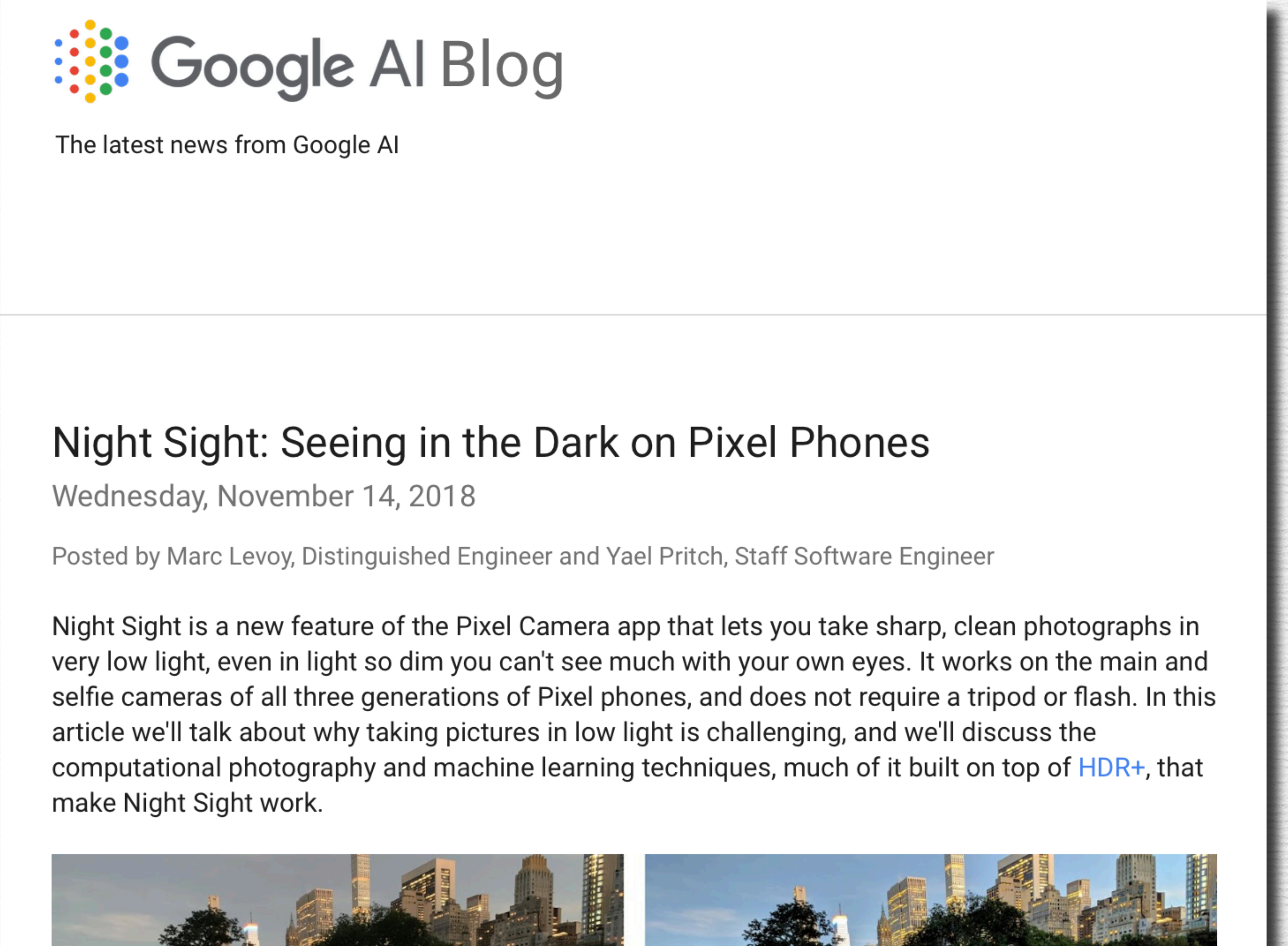
Results



- nearly as good as 2x optical zoom
- limited by diffraction spot size of lens

Example #4: Night Sight mode on Pixel 3

[Levoy and Pritch, Google AI blog]



Google AI Blog

The latest news from Google AI

Night Sight: Seeing in the Dark on Pixel Phones

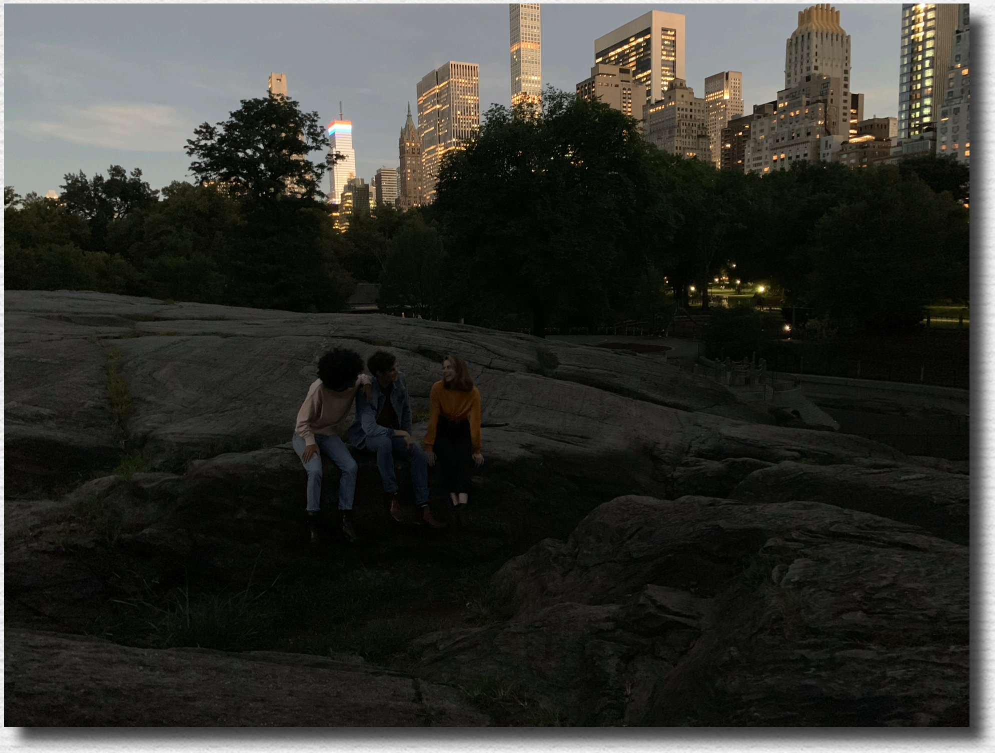
Wednesday, November 14, 2018

Posted by Marc Levoy, Distinguished Engineer and Yael Pritch, Staff Software Engineer

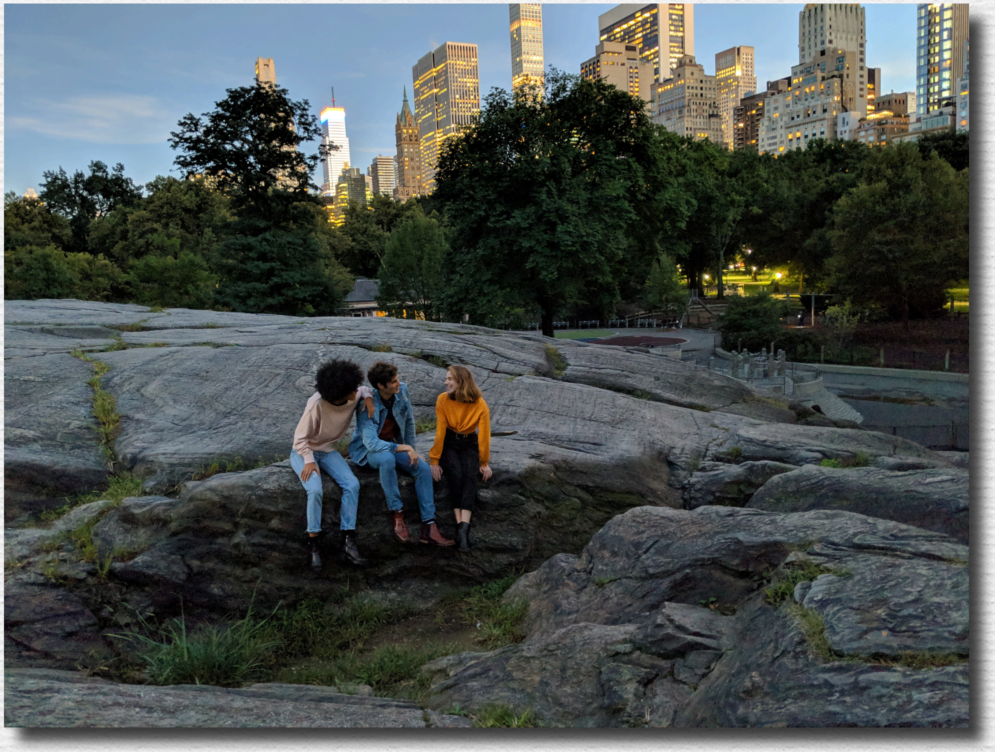
Night Sight is a new feature of the Pixel Camera app that lets you take sharp, clean photographs in very low light, even in light so dim you can't see much with your own eyes. It works on the main and selfie cameras of all three generations of Pixel phones, and does not require a tripod or flash. In this article we'll talk about why taking pictures in low light is challenging, and we'll discuss the computational photography and machine learning techniques, much of it built on top of HDR+, that make Night Sight work.

Example #4: Night Sight mode on Pixel 3

[Levoy and Pritch, Google AI blog]



iPhone XS

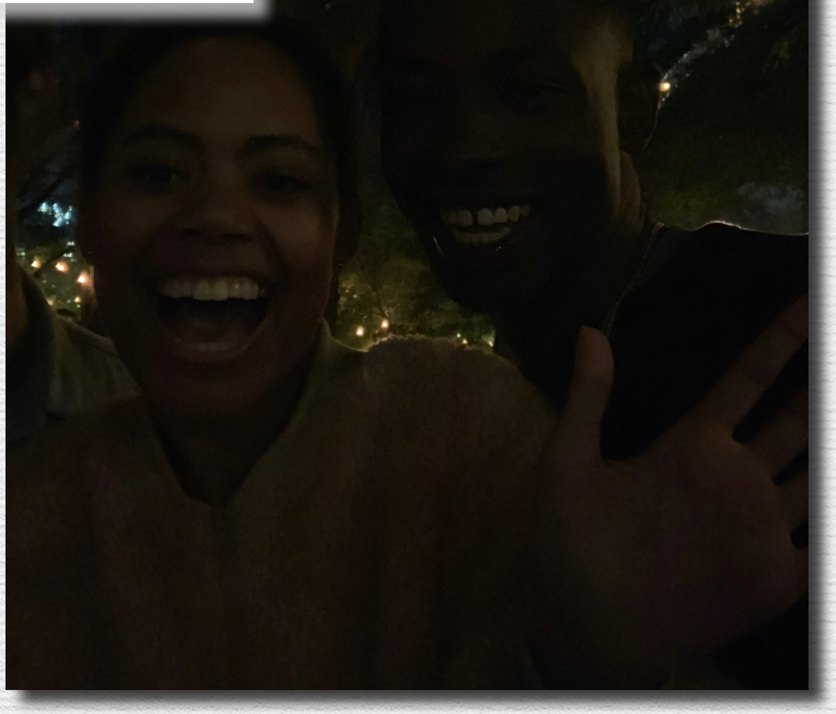


Pixel 3 with Night Sight

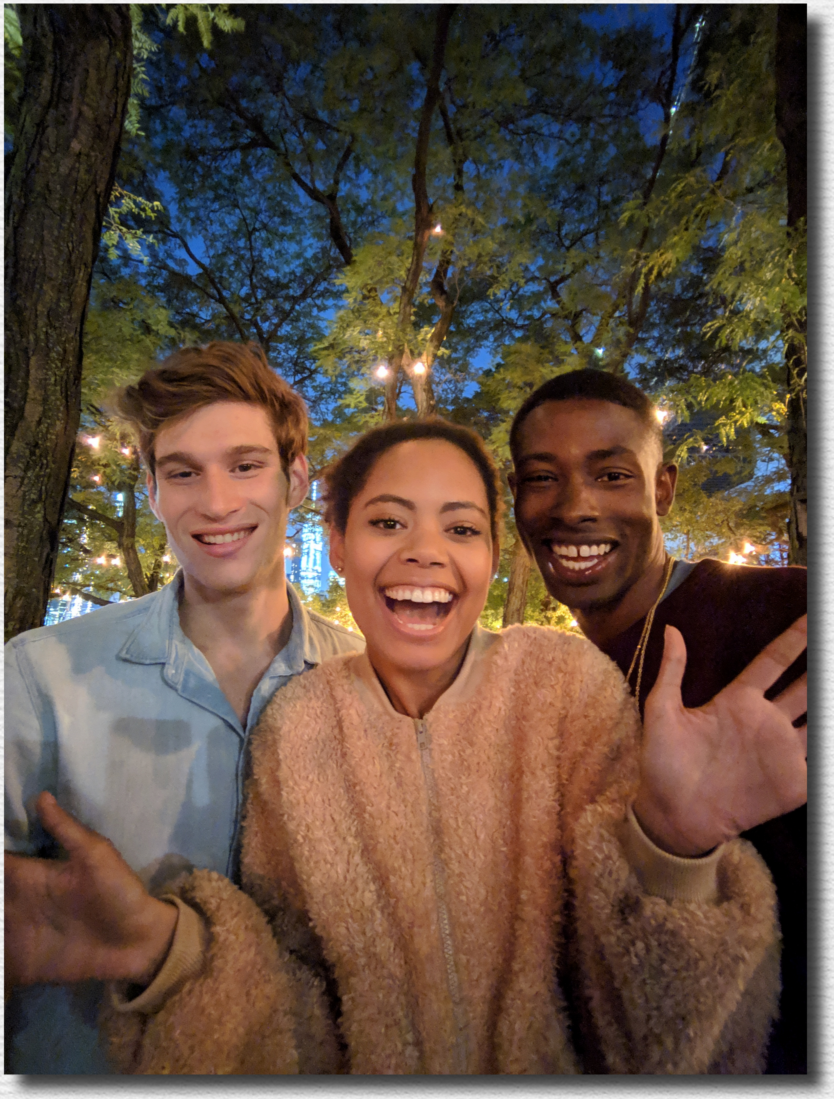
Example #4: Night Sight mode on Pixel 3

[Levoy and Pritch, Google AI blog]

(Synthcam and SeeInTheDark)



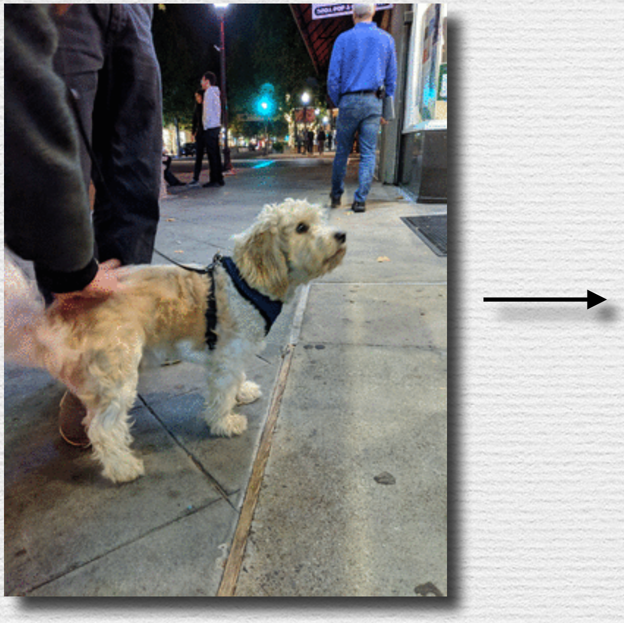
iPhone XS

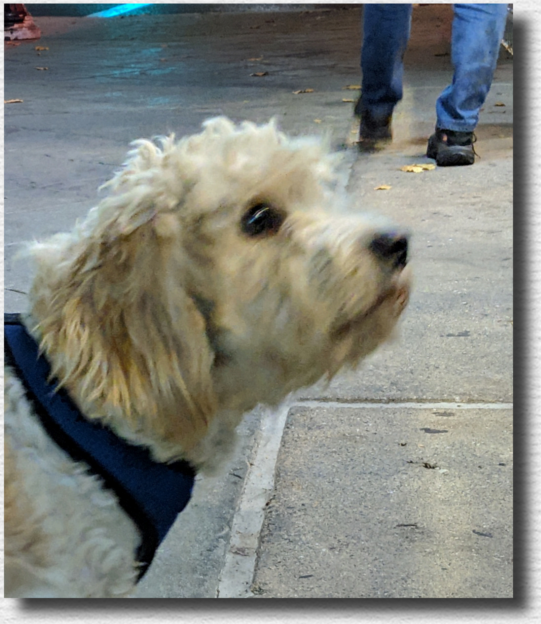


Pixel 3 with groupie camera and Night Sight

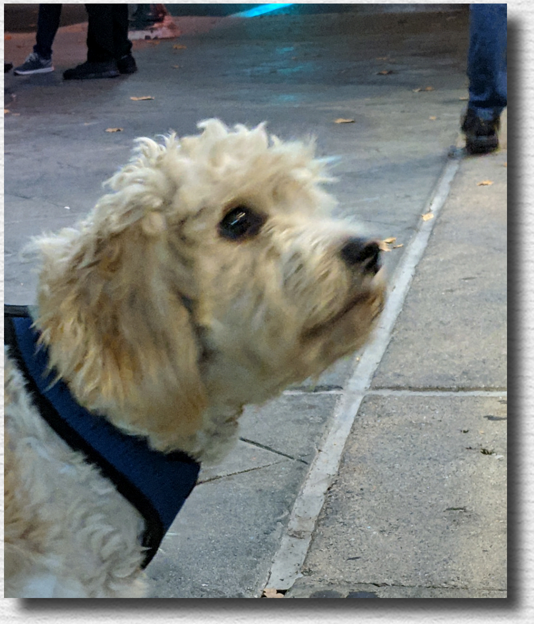
Technologies in Night Sight

- capture up to 15 frames after shutter press
  - animation telling user (and subject) to hold still
- motion metering
  - if handshake or moving objects, shorten each exposure





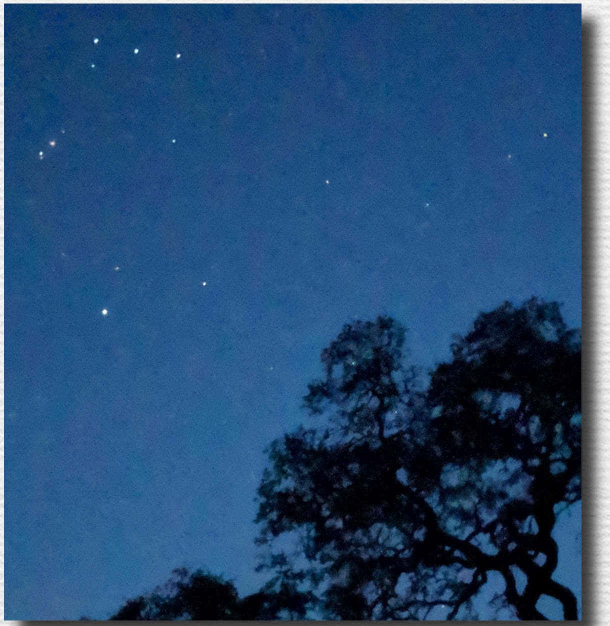
without motion metering



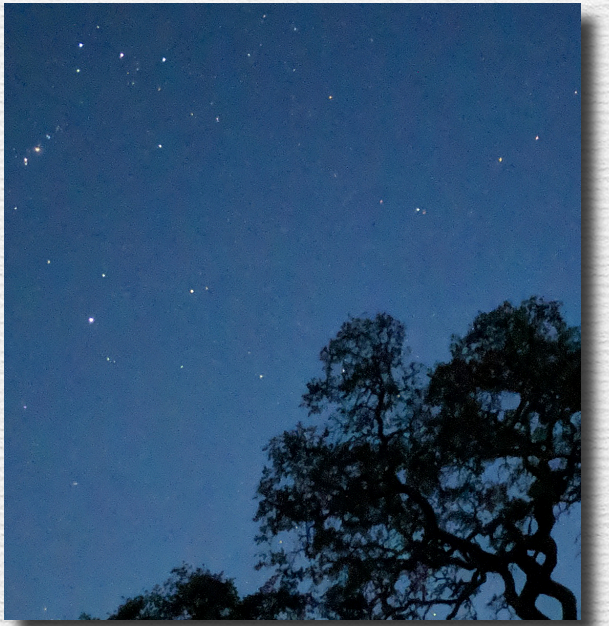
with motion metering

Technologies in Night Sight

- capture up to 15 frames after shutter press
  - animation telling user (and subject) to hold still
- motion metering
  - if handshake or moving objects, shorten each exposure
  - if on tripod, lengthen each exposure and total capture time



handheld



tripod

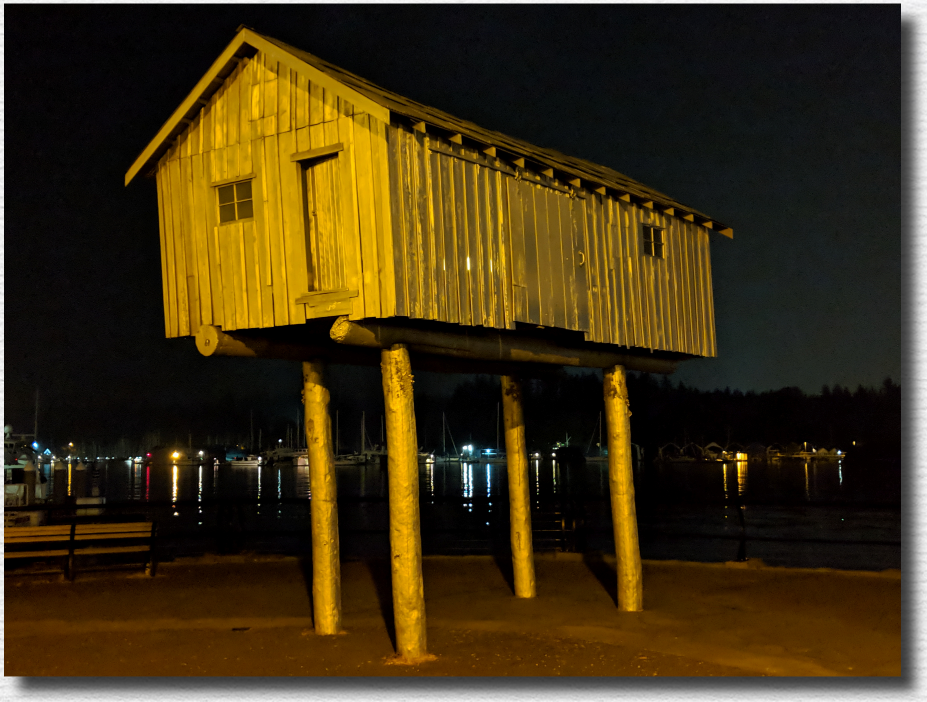
Technologies in Night Sight

- capture up to 15 frames after shutter press
  - animation telling user (and subject) to hold still
- motion metering
  - if handshake or moving objects, shorten each exposure
  - if on tripod, lengthen each exposure and total capture time
- robust align and merge
  - Super Res Zoom (Pixel 3)
  - HDR+ (Pixel 1 and 2)

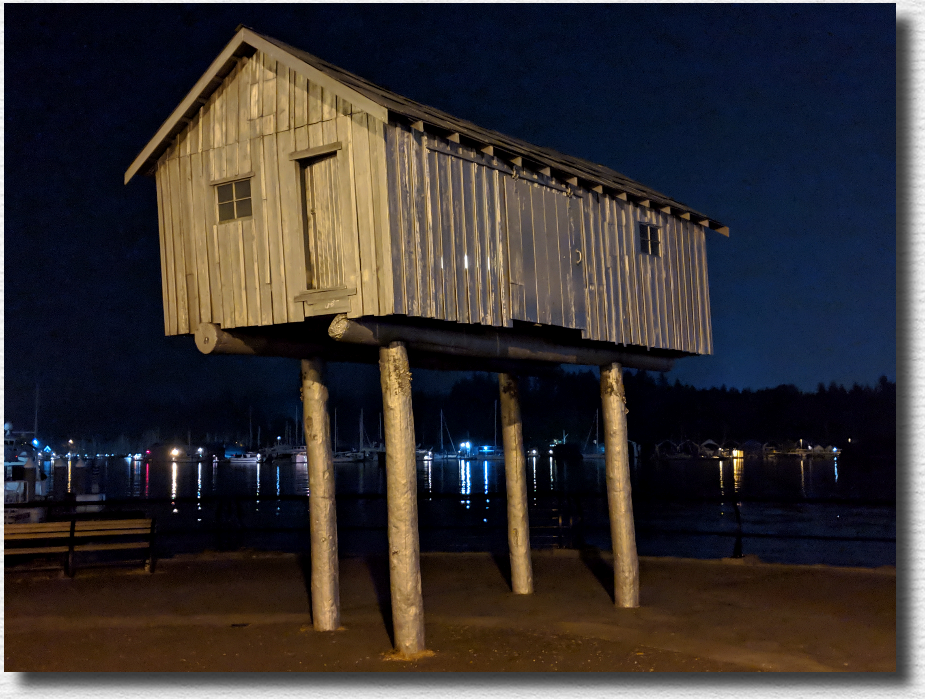
Technologies in Night Sight

- capture up to 15 frames after shutter press
  - animation telling user (and subject) to hold still
- motion metering
  - if handshake or moving objects, shorten each exposure
  - if on tripod, lengthen each exposure and total capture time
- robust align and merge
  - Super Res Zoom (Pixel 3)
  - HDR+ (Pixel 1 and 2)
- learning-based white balancing

Technologies in Night Sight



heuristics-based white balancer

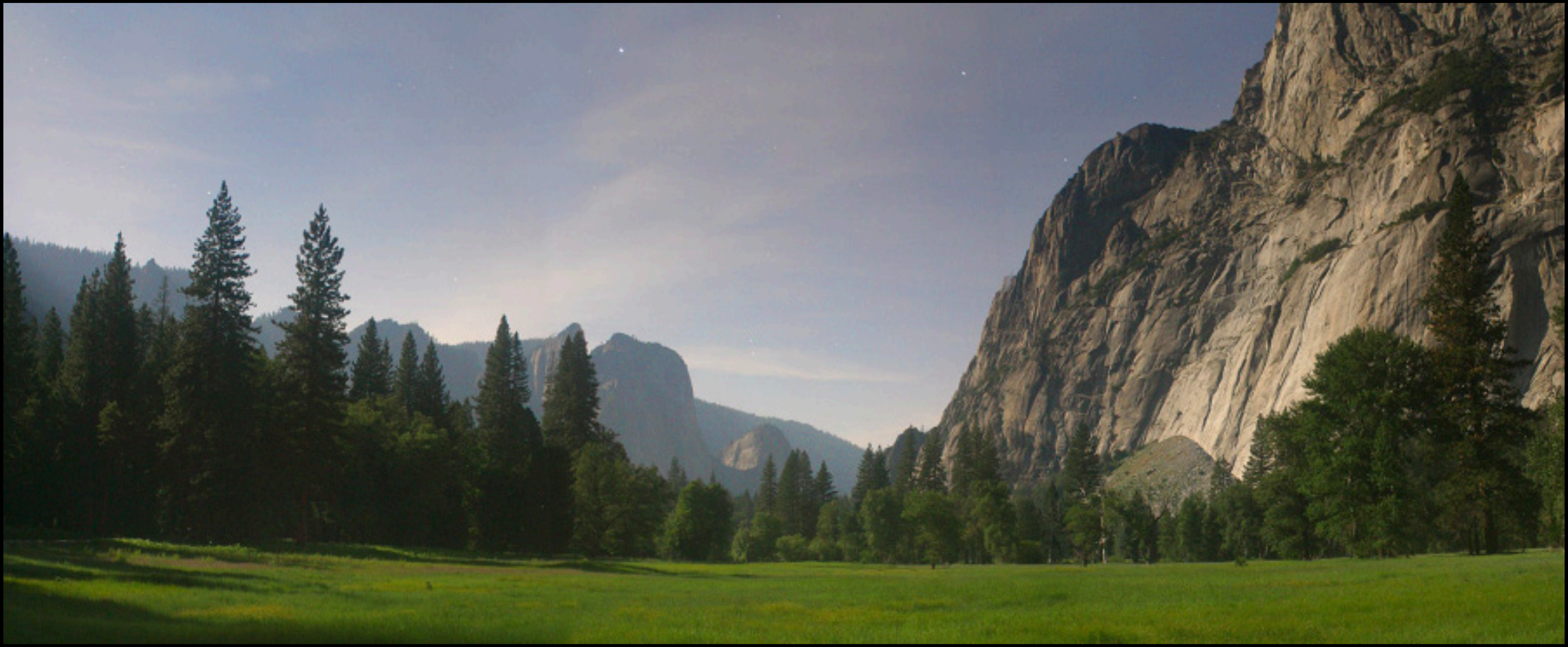


learning-based white balancer

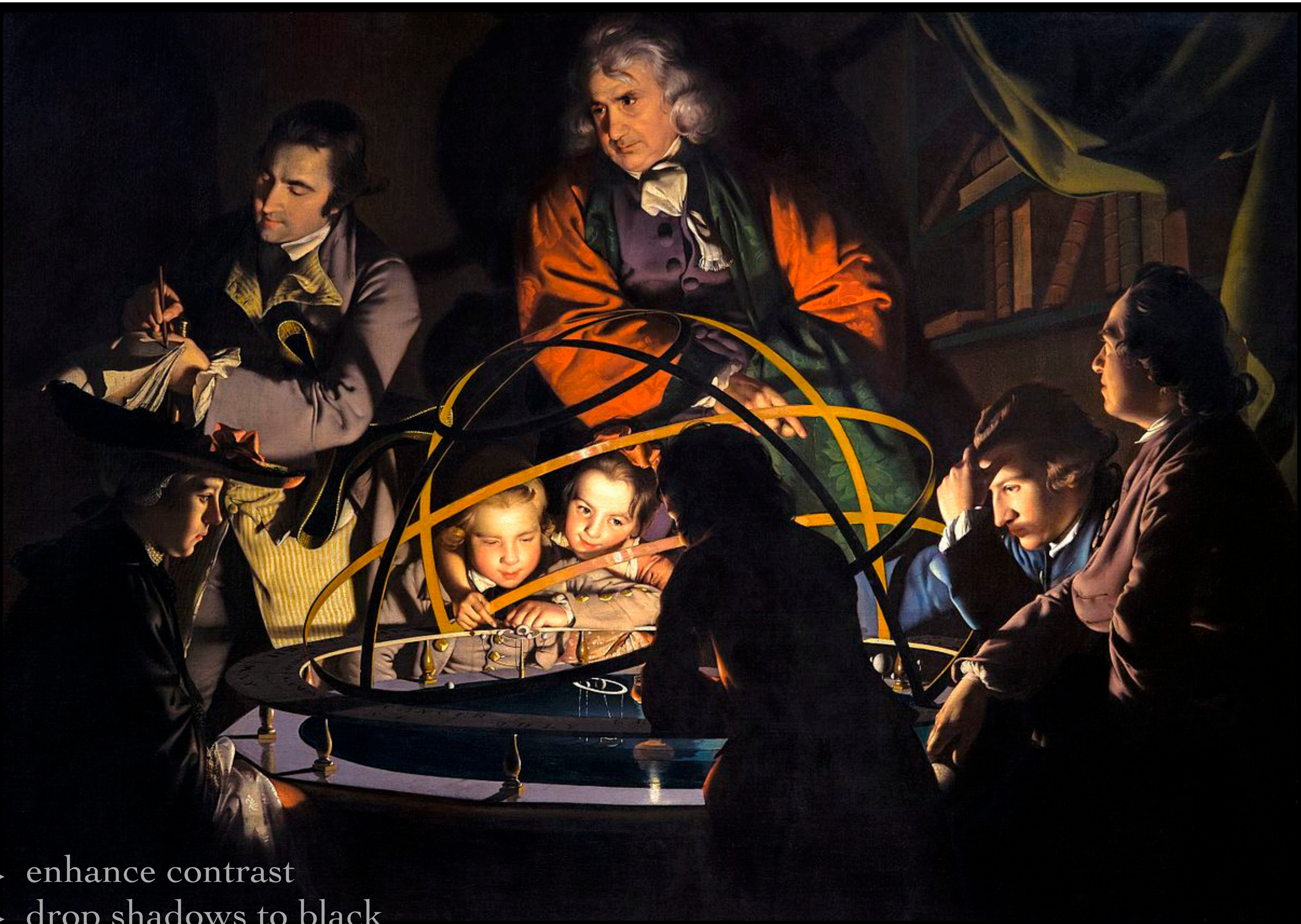
- learning-based white balancing

Technologies in Night Sight

- capture up to 15 frames after shutter press
  - animation telling user (and subject) to hold still
- motion metering
  - if handshake or moving objects, shorten each exposure
  - if on tripod, lengthen each exposure and total capture time
- robust align and merge
  - Super Res Zoom (Pixel 3)
  - HDR+ (Pixel 1 and 2)
- learning-based white balancing
- tone mapping to keep night looking like night



Jesse Levinson, Canon SLR, 24mm/1.4 lens, 3-minute exposure, ISO 100



- enhance contrast
- drop shadows to black
- surround scene with darkness

Joseph Wright of Derby, A Philosopher Lecturing on the Orrery (1766)

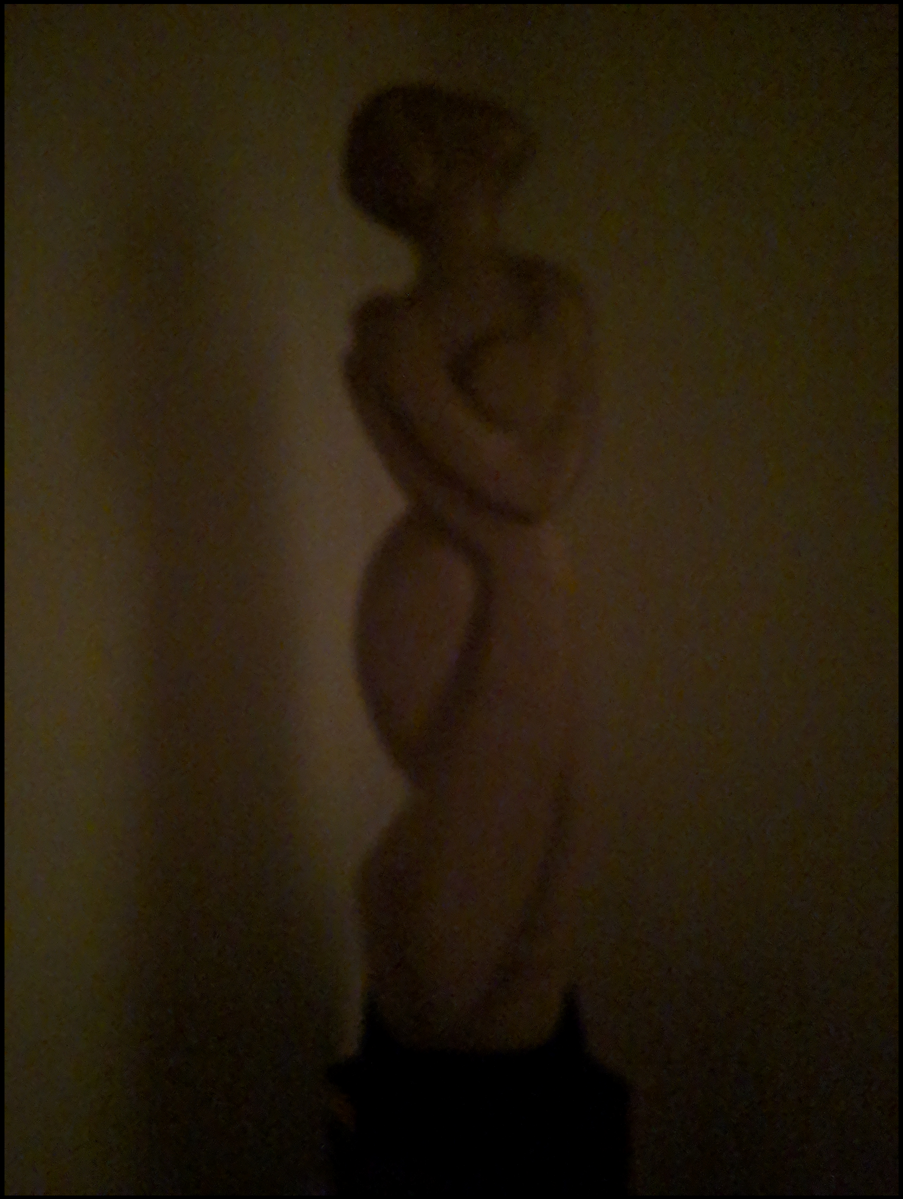
(Alex Savu, Redskull Productions)

Pixel 3 Night Sight

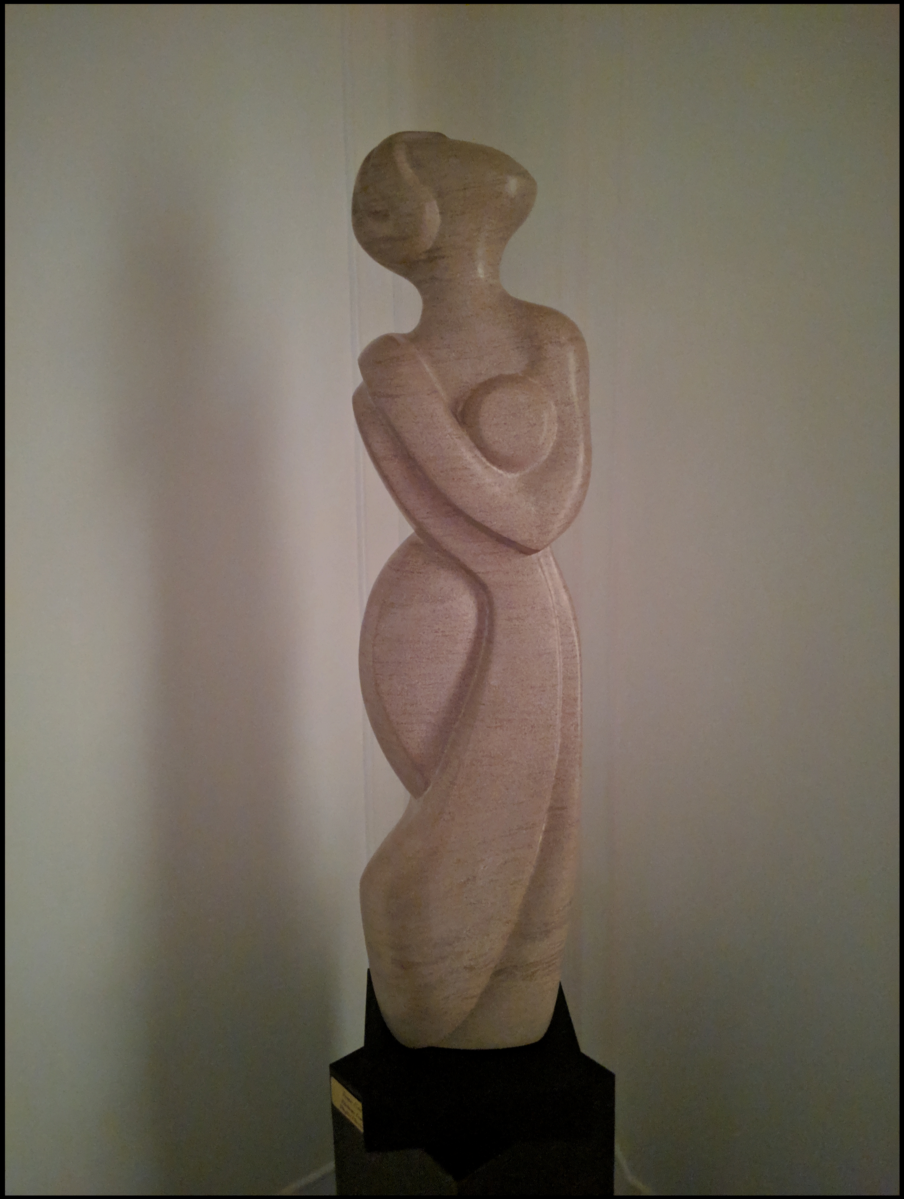
(Diego Perez)

(Sandeep Vijayasekar)

(Reed Bennett)



HDR+
(autofocus failed)



Night Sight
(handheld with manual focus)

(Marc Levoy)

Sculpture by Phyllis Chesler

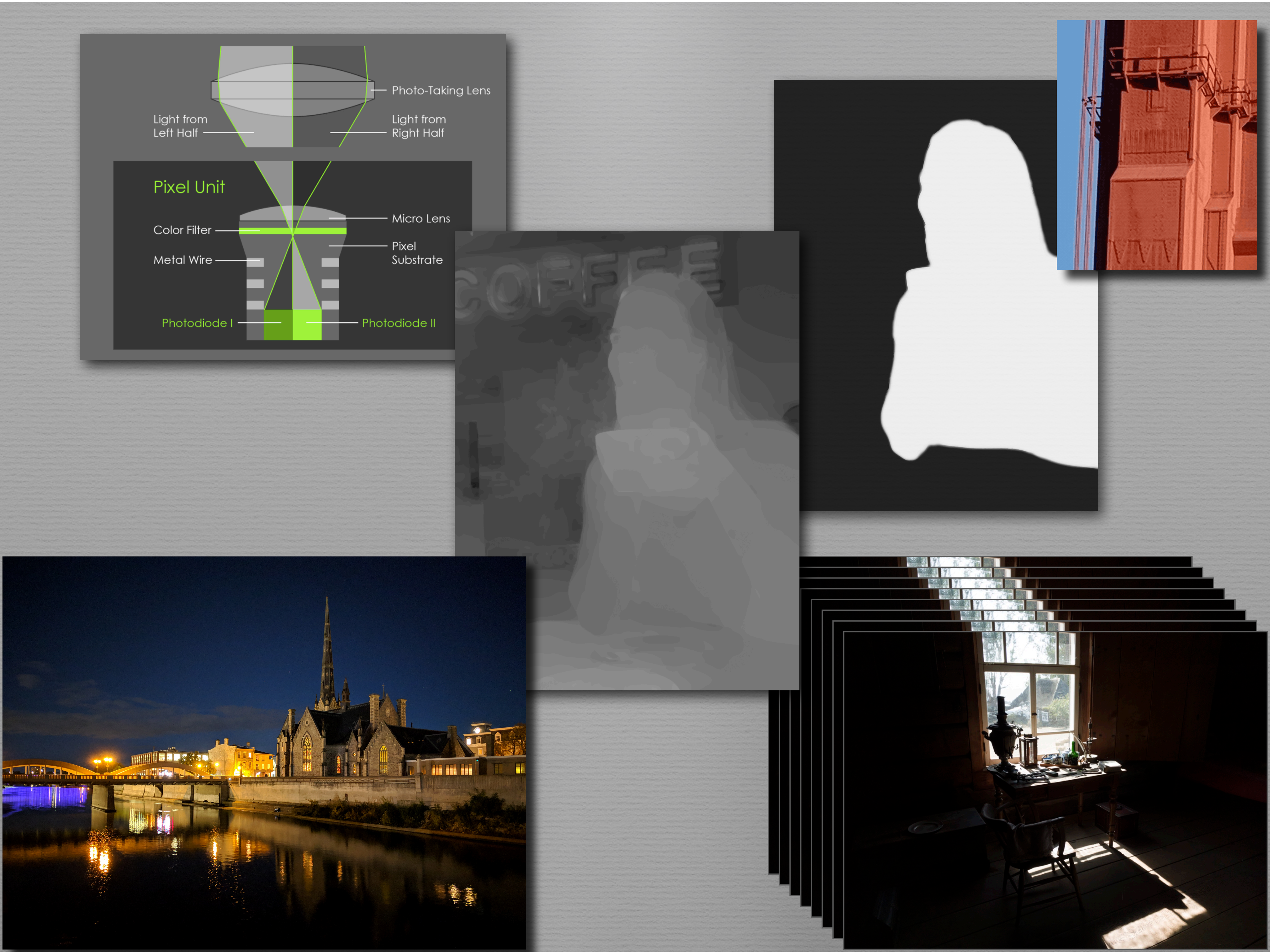


Photo-Taking Lens
Light from Left Half
Light from Right Half
Pixel Unit
Micro Lens
Color Filter
Pixel Substrate
Metal Wire
Photodiode I
Photodiode II
COFFEE